

McGRAW-HILL PUBLICATIONS IN PSYCHOLOGY

J. F. DASHIELL, PH.D., CONSULTING EDITOR

FUNDAMENTAL STATISTICS
IN PSYCHOLOGY AND EDUCATION

Fundamental Statistics

in

Psychology and Education

BY

J. P. GUILFORD

Professor of Psychology, University of Southern California

FIRST EDITION

FOURTH IMPRESSION

McGRAW-HILL BOOK COMPANY, INC.

NEW YORK AND LONDON

1942

Keep close to experience; add as little of your own as possible; if you have to add something, be mindful to give an account of every step you take.—F. M. URBAN.

is provided, and a novel, pedagogically simple derivation of the analysis-of-variance principle is presented. In a chapter quite new to this type of text, entitled Testing Hypotheses, much of Fisher's work is reflected, and chi square is given prominence. In another new chapter, entitled Predictions and Errors of Prediction, some new devices of practical importance are introduced. In this chapter and in others, much attention is given to enumeration data and the statistics of attributes, a field that is growing in importance in the social sciences in general. A treatment of factor analysis has been omitted for the reasons that this subject cannot any longer be adequately presented in the space that a text of these proportions would permit, and its study and mastery extend well beyond the student's first year of statistics. Even the final chapter on Mental Tests had to be treated rather sketchily in order to remain within reasonable bounds of space allotted to the volume. In general, there was recognition of the limitations, self-imposed, where references to more advanced treatments were regretfully made in order to stay within the bounds of a fundamental statistics.

The author gladly expresses acknowledgments and thanks to Prof. Harry Helson, who read and criticized three of the chapters. To H. M. Cox, with whom the author was associated in the Bureau of Instructional Research, he owes much for certain ideas regarding ways of presentation of data and concerning the selection of useful methods. To his wife, Ruth B. Guilford, the author is, as always, most indebted for constant help in the preparation of the manuscript. To publishers and authors who have generously permitted the reproduction or use of material he is grateful. These and other contributions are acknowledged specifically at various places in the volume. To Prof. R. A. Fisher and to Messrs. Oliver & Boyd of Edinburgh the author is indebted for permission to reprint Table E from their book *Statistical Methods for Research Workers*, 8th ed., 1942.

J. P. GUILFORD.

SANTA ANA, CALIF.,
September, 1942.

CHAPTER	PAGE
VII SOME APPLICATIONS OF THE NORMAL CURVE	95
The Use of Standard Scores	96
The <i>T</i> -scale and <i>T</i> -scaling of Tests	99
The <i>C</i> -scale System	104
Measurements from Judgments of Rank Order	107
Scaling Judgments in Absolute Categories	110
Scaling Test Items for Difficulty	114
Transforming One Distribution into Terms of Another	118
Exercises	122
VIII THE RELIABILITY AND SIGNIFICANCE OF STATISTICS	125
The Reliability of Averages	125
The Reliability of Other Statistics	132
The Reliability of Differences	135
Analysis of Variance	145
Exercises	153
IX. TESTING HYPOTHESES	156
The Null Hypothesis	156
Chi Square	167
Exercises	173
X. PREDICTION AND ERRORS OF PREDICTION	176
Predicting Attributes from Other Attributes	178
Predicting an Attribute from Measurements	181
Predicting Measurements from Attributes	188
Predicting Measurements from Other Measurements	192
Exercises	196
XI. CORRELATION METHODS	198
How to Compute a Coefficient of Correlation	202
Regression Equations	211
Interpretations of a Coefficient of Correlation	218
Assumptions Underlying the Product-moment Correlation	223
Exercises	224
XII. OTHER CORRELATION METHODS	227
Spearman's Rank-difference Correlation Method	227
The Correlation Ratio	231
The Biserial Coefficient of Correlation	237
Tetrachoric Correlation	240
The Phi Coefficient	245
Some Special Problems in Correlation	248
Exercises	254
XIII MULTIPLE AND PARTIAL CORRELATION	256
Multiple Correlation	256
Partial Correlation	268
Exercises	271

task. In memory experiments, we measure learning efficiency in terms of the number of trials to attain a certain standard of performance or in terms of the "goodness" of performance at the end of a certain trial or time. We measure efficiency of retention in terms of the time required for relearning (overcoming the forgetting that has taken place) and the efficiency of recall in terms of association time or in terms of the number of items correctly recited.

In the sphere of motivation, we gauge the strength of drive in terms of the amount of punishment (electric shock) an organism (for example, a rat) will endure in order to reach his immediate goal or in terms of the number of times he will take a constant punishment in order to attain the same result. The difficulty of a task or test item can now be specified in quantitative terms, as can the affective value (degree of liking or disliking) for a color, a sound, or a pictorial design. In studies of sensory and perceptual powers, the threshold stimulus and the differential limen are given in terms of stimulus magnitudes. The span of perception or of apprehension is given in terms of the average number of items that the observer can report correctly after momentary exposures. The galvanic skin response, the pupillary response, and the amount of salivation also serve as quantitative indicators of amounts of psychological happenings.

Some Examples of Educational Measurement.—Many an educational problem is also a psychological problem, and its mode of measurement has been indicated in the preceding paragraph. Achievement in any area of learning, like any mental ability, is measurable in terms of test scores. Marks, however obtained, have been the traditional mode of evaluating students in specific units of formal education. Attendance records, data on size of classes, on budgets, on supplies, and on other material aspects of the well-regulated school system compose another list of measurements in education. Outcomes of educational effort are often expressed quantitatively in terms of promotion statistics, achievement ratios, and estimates on teaching success. Whether for purposes of research in education or for systematic and meaningful record keeping, statistical methods become an indispensable kind of tool.

Some Different Kinds of Measurement.—In a superficial way, it is easy to see, as one glances over the list of psychological and educational measurements just mentioned, that there are different kinds of measurement involved. Among the psychologist's measurements, some are in terms of the stimulus—for example, the threshold stimulus or stimulus difference; the number of syllables or items; the amount of electric

an unrivaled device for bringing order out of chaos; of seeing the general picture in one's results.

4. *They enable us to draw general conclusions*, and the process of extracting conclusions is carried out according to accepted rules. Furthermore, by means of statistical steps, we can say about how much faith should be placed in any conclusion and about how far we may extend our generalization.

5. *They enable us to make predictions* of "how much" of a thing will happen under conditions we know and have measured. For example, we can predict the probable mark a freshman will earn in college algebra if we know his score in a general scholastic-ability test, his score in a special algebra-aptitude test, his average mark in high-school mathematics, and perhaps the number of hours per week that he devotes to studying algebra. Our prediction may be somewhat in error because of other factors that we have not accounted for, but our statistical methods will also tell us about how much margin of error to allow in our predictions. Thus not only can we make predictions but we know how much faith to place in them.

6. *They enable us to analyze some of the causal factors out of complex and otherwise bewildering events.*—It is generally true in the social sciences and in psychology and education in common with them that any event or outcome is a resultant of numerous causal factors. The reasons why a man fails in his business or in his profession, for example, are varied and many. Causal factors are usually best uncovered and proved by means of experimental method. If it could be shown that, all other factors being held constant, certain business men fail to the extent that they possess some defect of personality "X," then it is probable that X is a cause of failure in this type of business. Unfortunately for the social scientist, he cannot manage men and their affairs sufficiently to set up a good experiment of this type. The next best thing is to make a statistical study, taking business men as we find them, working under conditions as they normally do. The life-insurance expert does the same kind of thing when he follows the trail of all possible factors that influence the length of life and determines how important they are. On the basis of these statistical findings, he can predict about how long an individual of a certain type will probably live, and his insurance company can plan his insurance policy accordingly. Statistical methods are therefore often a necessary substitute for experiments. Even where experiments are possible, the experimental data must ordinarily receive appropriate statistical treatment. Statistical methods are hence the constant companions of experiments.

not an absolute zero point is the centigrade thermometer. The zero point is arbitrarily placed at the freezing point of water. With this instrument, we can say that the temperature of the weather changes as much when it rises from 0 to 25 as it does when it rises from 25 to 50. But we cannot say that 50° is twice as warm as 25° or that 100° is twice as hot as 50° . We can thus add and subtract measurements on this scale and get sensible answers, but we cannot multiply and divide. If we translate our zero mark to the absolute zero point (zero heat), which, in terms of the common thermometer is -273° , then we can perform these operations. On the absolute scale, our 25° becomes 298° , and our 50° becomes 323° . Now it is obvious that the higher of the two (323) is not two times the lower (298). But if our absolute centigrade scale is correct, with regard to equality of units, we may well say that a temperature twice as hot physically as 298° is a temperature of 596° (also on the absolute scale).

Mental-test Scales as Metric Devices.—What shall we say of a measuring scale of the type most frequently used in psychology and education—mental-test scores in terms of number of items correct? Have we here a scale with absolute zero and equal units? Usually not. A score of zero, no items correctly answered, does not mean zero ability. For had we included some easier items, even the lowest individual in the test could probably have made a score numerically greater than zero. Thus we are unable to say that a score of 50 points means twice the ability represented by a score of 25 or half the ability represented by a score of 100 points. For if our real zero-ability score should have been some 25 points below our arbitrary one, these three scores would then become 50, 75, and 125.

Now the second is *not* twice the first or half the third. Nor can we be sure that our units are equal within the range of scores obtained. Unless the units were equal, we should not be able to say that a score of 100 is as far above one of 75 as the latter is above a score of 50. As a matter of long experience, however, we find that test scores generally behave as if units were equal; as if one item correct adds an amount to the measurement of ability equal to that added by any other item correct. There are various indications that tell the experienced worker in statistics when his measurements probably possess equal units and when they do not. And when they do, we can proceed to apply most of the ordinary statistical procedures. When we strongly suspect that they do not, we can make adjustments or substitute other statistical methods that do apply. The beginner in statistical work

discrete measurements, as, for example, the number of children in a family, we customarily proceed *as if* 8 children meant anywhere from 7.5 to 8.5. The only notable exception to this general rule is in dealing with chronological age as given to the *last* birthday and the like. Then a twelve-year-old child is anywhere from 12.0 to 13.0. If ages are given *to the nearest birthday*, however, our rule again applies, and a twelve-year-old falls in the interval 11.5 to 12.5.

SOME RULES REGARDING NUMBERS

Approximate and Exact Numbers.—Measurements, when taken to the nearest whole unit, are known as *approximate numbers*. They are always “fuzzy” and are of uncertain value within the unit where they fall. When we find a number by enumeration of discrete objects, we have exact numbers; for example, 15 men, 42 letters, or 50 pencils. The distinction between exact and approximate numbers we shall find very important when they are used in calculations. Some important rules about calculations are presented next. They would be unnecessary if all numbers in statistics were exact.

How to Round Numbers.—The beginner in statistical computation invariably asks, “How many decimal places shall I save?” In just this form, the question cannot be answered. The question should read instead, “How many significant figures have I?” A number may have been rounded, dropping *all* digits to the right of the decimal point, yet not all of the remaining figures may be significant or exact. Another number may have four places remaining to the right of the decimal point, yet all of them may be significant. Some students may, if they lack good rules, drop too many figures, thus losing much of the accuracy that they really have; others may save a string of figures beyond the limit of significance, giving the appearance of great accuracy that is really fictitious.

First let us be clear as to the proper way to round a number. There is no particular difficulty in rounding to the nearest whole number; 15.7 becomes 16, and 27.4 becomes 27; 9.6 becomes 10, and 0.96 becomes 1. In rounding to two decimal places, the same principles apply; 2.1827 becomes 2.18, and 91.2179 becomes 91.22. It is when the first digit to be dropped is 5 that difficulties arise. In rounding to two decimal places, again, the number 7.1654 becomes 7.17, and even 7.16502 becomes 7.17 rather than 7.16, for the reason that the decimal fraction beyond the 6 is greater than just .00500. Had the number been 7.16499, we should have rounded to 7.16, because it is a shade closer to 7.16 than to 7.17.

digit fixes the number between .4195 and .4205. A lone zero before the decimal point, however, as 0.41, is not significant, since it adds nothing to our information concerning numerical value.

Rules Governing Significant Figures in Computation.—The following rules will determine how many significant figures there are in a number found by computation.

1. *In Sums of Numbers.* CASE I.—When all the numbers added are correct to the nearest unit, the sum is regarded as correct to the nearest unit.

Example: $47 + 161 + 5,171 = 5,379$, a sum correct to four significant figures

A similar case occurs when all the numbers added have the same number of decimal places.

Example: $2.91 + 40.22 + 0.07 = 43.20$, where the answer is significant to the second decimal place because all the numbers were significant to that place

CASE II.—When numbers that are not correct to the same number of places at the right of the decimal point are added, the sum is significant only as far as the number having the *smallest* number of decimal places.

Example: $17,257 + 142.1 + 75.47 = 234.8$, which is rounded from 234.827. Note that the rounding was done *after* summing and not before.

A similar situation is true when numbers rounded to the *left* of the decimal point are summed.

Example: $75,000 + 3,845 = 79,000$, which is rounded from 78,845 because in the first number there are only two significant digits to the left of the hundreds place

2. *In Differences.* CASE I.—If the two numbers are significant to the same digit at the right, the difference is also significant that far to the right.

Example: $173.24 - 94.84 = 78.40$, the zero being significant.

Frequently a difference is drastically reduced in the number of significant figures, so much so that further computations with this difference are sometimes lacking in desired accuracy. This situation is to be avoided when possible.

Example: $4.692 - 4.685 = 0.007$

CASE II.—As with addition, the answer is significant no further to the right than is the least significant number.

there are not three pairs, but there are three groups, one of them being an incomplete pair; so there will be three significant digits in the square root, or 16.9. Again, the number 4451.927 divides up (starting at the decimal point) as 44 51.92 7, and its root has four significant figures, which are 66.72.

CASE II.—The square root of an exact number may be given to as many places as one wishes.

Example. $\sqrt{5} = 2.2361$ This could be carried further, or we could round it to 2.236 or to 2.24, depending upon our purposes

Some Exceptions to the Rules.—Although the rules as just given are acceptable and sound, there are times when we should properly depart from them. One frequently has to use his best judgment and do the most reasonable thing. To follow the rules rigidly at every step of the way would sometimes introduce inaccuracies or else cause one to lose information that he really has and needs. One good general principle to follow is to *carry along more than the recognized significant figures through the successive steps of calculation and withhold the rounding of numbers until the final answer is obtained*, such as an arithmetic mean, a standard deviation, or a correlation coefficient. Further suggestions will be offered more appropriately later when we are dealing with specific cases.

Exercises

1. State the exact limits to the following scores or measurements. 57 sec. 150 kg. 65 score points 0 score points 14 5 cm .125 sec. 15 years (to the last birthday)
2. Round the following numbers to one decimal place: 26 418 4 072 4.98 9 092 120 052 0 3500 44 7508 291 6500 8.8502 31 15— 48 25+
3. How many significant figures in each of the following numbers: 1,942 20,007 170 9 0 31 28,000 0.0017 0 3400 21,500
4. Write the answers to the following problems to as many significant figures as the rules allow:
 - a. 21.3 in. times 15 (where 15 is an exact number).
 - b. $5.2 + 17.2509 + 918.04$
 - c. 242.8×0.075 .
 - d. 4 27505 divided by 25 (where 25 is an exact number).
 - e. 17.98 divided by 2 1.
 - f. 38.6 squared.
 - g. $\sqrt{50}$ (where 50 is an exact number, but be reasonable)
 - h. $\sqrt{25.3179}$.

Some Suggestions to the Student

A Review of Arithmetic and Elementary Algebra.—Some students who have not kept alive the skills they once acquired in arithmetic and elementary algebra fre-

CHAPTER II

FREQUENCY DISTRIBUTIONS

After we obtain a set of measurements, the next customary step is to put them in systematic order by grouping them in classes. A set of individual measurements, taken as they come, as in the list in Table 1,

TABLE 1—SCORES IN AN INK-BLOT TEST

25	33	35	37	55	27	40	33	39	28
34	29	44	36	22	51	29	21	28	29
33	42	15	36	41	20	25	38	47	32
15	27	27	33	46	10	16	34	18	14
36	21	19	26	19	17	24	21	27	16

does not convey much useful information to us. We have merely a vague, general conception of about how large they run numerically but that is about all. The data in Table 1 are scores made by 50 students in an ink-blot test. Each score is the number of objects the student reported in observing 10 ink blots during a period of 10 min. Concerning such a set of data we usually want to know several things. One is what kind of score the average or typical student makes; another concerns the amount of variability there is in the group or how large the individual differences are; and a third is something about the shape of the distribution of scores, *i.e.*, whether the students tend to bunch up at either end of the range or at the middle or whether they are about equally scattered over the entire range. The first steps in the direction of answering these questions require the setting up of a frequency distribution.

THE CLASS INTERVAL—ITS LIMITS AND FREQUENCIES

The Size of Class Interval.—We could begin by asking how many scores of 25 there are, of 26, 27, etc., but this would not give us an adequate picture, because in a group of only 50 individuals whose scores range from 10 to 55, many scores do not occur at all or are not repeated. We therefore combine the scores into a relatively small number of *class intervals*, each class interval covering the same range of scores on the scale of measurement.

and 14; in the next higher interval, scores of 15, 16, 17, 18, and 19; etc. Instead of writing out all the scores for each interval, we give only the bottom and top scores. Our intervals are then labeled 10 to 14, 15 to 19, 20 to 24, etc., or, more often, 10-14, 15-19, 20-24. Note that the bottom and top scores are given, and they represent what we call the *score limits* of the interval. They do not indicate exactly where each interval begins and ends on the scale of measurement. The score limits are useful primarily in tallying and in labeling the intervals.

Exact Limits of Class Intervals.—We shall soon find that in computations we must think in terms of *exact limits*. Remember that a score of 10 actually means from 9.5 to 10.5, and that a score of 14 actually means from 13.5 to 14.5. This means that the interval containing scores 10 to 14 inclusive actually extends from 9.5 to 14.5 on the measurement scale. Likewise, the interval having score limits of 15 and 19 has exact limits on the scale of 14.5 and 19.5. The interval labeled 55 to 59 actually extends from 54.5 to 59.5. The same principle holds no matter what the size of interval or where it begins. An interval labeled 14 to 16 includes scores 14, 15, and 16 and extends exactly from 13.5 to 16.5. An interval labeled 70 to 79 extends from 69.5 to 79.5. It will be seen that by following this principle each interval begins exactly where the one below leaves off, which is as it should be.

Tallying the Frequencies.—Having decided upon the size of class interval and with what scores to start the intervals, we are ready to list them, as in Table 2. It is accepted custom to place the highest measurements at the top of the list and the lowest at the bottom, as shown here. Space is left in the second column for the tallying process. Taking each score in Table 1 as we come to it, we locate it within its proper interval and write a tally mark in the row for that interval. Having completed the tallying, we count up the number of tally marks in each row to find the *frequency* (f) or total number of individuals falling within each group. The frequencies are listed in the third column of Table 2.

Checking the Tallying.—Next we sum the frequencies, and if our tallying has omitted none and duplicated none, the sum should equal the number of individuals. At the bottom of the column we find the symbol Σf , in which Σ (capital Greek sigma) stands for "the sum of" and so Σf is "the sum of the frequencies." The total number of individuals or measurements in our sample is symbolized by the capital letter N , which stands for "number." If Σf does not equal N , there has been a mistake in tallying, and tallying should be repeated until

is more accurately represented and the numbers of cases in each interval are more exactly shown. Figure 1 is of the type known as *frequency polygon*, and Fig. 2 is of the type called *histogram*, or sometimes, though less often, *column diagram*.

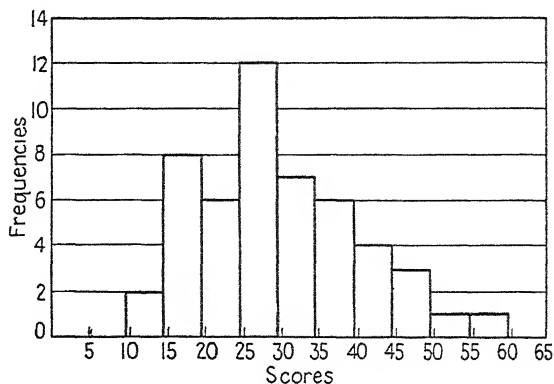


FIG. 2.—A histogram for the same distribution as in Fig. 1.

The Frequency Polygon and How to Plot It.—A polygon is a many-sided figure, and so the picture in Fig. 1 derives its name. There are a number of factors to be kept in mind in drawing such a figure.

The Kind of Graph Paper.—First, it might be said that, in general, the most convenient type of cross-section paper is the type that is ruled into heavy lines 1 in. apart each way, subdivided into tenths of an inch more lightly drawn.

The Width of the Diagram.—Second, the question of the height and width of the entire figure comes up. For the sake of easy readability, the width of the figure on this kind of graph paper should be at least 5 in. We have altogether 10 class intervals in which there are frequencies, but in drawing the diagram, we should allow for one more class interval at each end of the scale, making 12 in all. This is to permit bringing the ends of the polygon down to the base line (see Fig. 1).

Labeling the Base Line.—In deciding how many intervals to allow to the inch, it is well to remember that we are going to label the base line of the figure in terms of our measuring scale and hence should plan things so that $\frac{1}{10}$ in. will stand for an integral number of units on this original scale. In the ink-blot data, we have been dealing with a class interval of 5 units, and we are making room for 12 intervals on our base line—in other words, for 60 units. By allowing $\frac{1}{10}$ in. to each unit ($\frac{1}{2}$ in. to each class interval), our distribution will spread

the upper limit, 14.5 minus 2.5, and you also have exactly 12 as the midpoint, or the average of 10 and 14 is 12. The midpoint of the interval 55–59 is 57. When the class interval is 5 and the lowest score in each interval is a multiple of 5, as will be true in many of the instances met in psychology and education, the midpoints will end in 2 and 7 systematically. For the sake of a complete picture of the midpoints for the data in Table 1, we have given in Table 3 the full set of midpoints.

Plotting the Points.—Having determined the midpoints and knowing the frequencies corresponding to them, we are ready to plot the dots for the frequency polygon. For the two intervals at the ends of the distribution (see Table 3) we have frequencies of zero. Sometimes there are frequencies of zero *not* in the last two classes. When so, we plot these dots also on the base line and bring the lines that connect the dots down to the base line at those places. That did not happen to be the case in these data. When the dots are placed at the midpoints, as directed, it may be noted that they do not appear directly above the midpoints of the marked places on the base line (5, 10, 51, 20, etc., in this case). Remember that these multiples of 5 are *not* the exact limits of the class intervals; they are merely convenient and meaningful reference points on our original scale. Had we begun the class intervals at scores other than multiples of 5—for example, at 11, 16, 21, 26, etc.—we should still plot at the midpoints of the intervals (now different than before) and should still label the reference points as multiples of 5, as in Fig. 1. The curve as drawn truly represents the shape of the distribution as we have grouped the scores.

The Histogram and How to Plot It.—Many of the facts learned in plotting the frequency polygon also apply in plotting the histogram. The choice of size, proportions, units per square of graph paper all are the same. The only important difference is that although we locate the height of each column or rectangle by placing a dot at the midpoint of each interval, we do not then connect dot to dot with straight diagonal lines. Instead, we draw a short horizontal line through each dot (see Fig. 2), extending it to the upper and lower *exact* limits of each class interval. Those exact limits are given in Table 3 for our data. Having done this, we erect vertical lines at each of these exact limits tall enough to form complete rectangles. Again it may be noticed that the rectangles seem to be misplaced a half unit with respect to the numbers on the base line, but this is correct; the choice of limits for our classes makes the exact limits come a half unit below the multiples of 5, *i.e.*, at 4.5, 9.5, 14.5, 19.5, etc.

figure. If the smaller distribution is large enough to be clearly legible, the larger one may extend beyond reasonable bounds. Furthermore, if it is general shape and general position on the measuring scale and dispersion that we wish to compare, the marked difference in size may make such comparisons very unsatisfactory. A common solution to this difficulty is to reduce both distributions to *percentage frequencies* instead of plotting the original frequencies. It is then as if we had two distributions whose N 's equal 100. This makes their areas approximately equal in the polygon form, and comparisons of shape, level, and dispersion are then quite satisfactory.

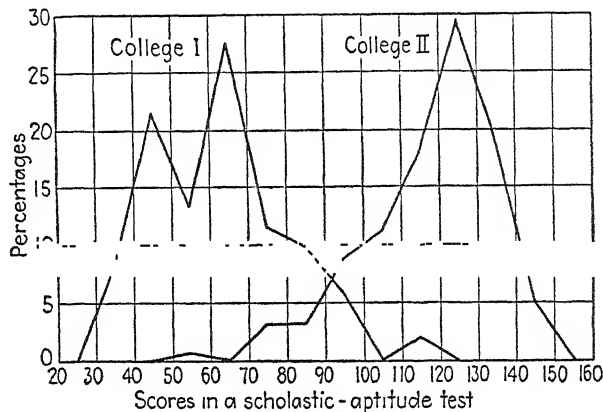


FIG. 3—Distributions of scores in an aptitude test in two colleges. Frequencies have been reduced to a percentage basis.

How to Find Percentage Frequencies.—As an example of how to transform frequencies into percentages the data in Table 4 are presented. In each case, the frequencies in the distribution are multiplied by 100, then divided by N . A shorter procedure would be to find the quotient $100/N$ to four or more decimal places (both are exact numbers), then multiply each frequency in turn by this ratio. In distribution I, the ratio is $100/51$, which equals 1.960784, and in distribution II it is $100/160$, which equals 0.625. Multiplying each frequency f_1 by 1.960784, we obtain the list of percentages in column (4), and multiplying each frequency f_2 by 0.625, we obtain the list in column (5). Plotting these percentages above the corresponding midpoints of class intervals, we obtain the distribution curves in Fig. 3. Although it was apparent in Table 4 that the second group were higher on the scale than the first and that there was still considerable overlapping of scores between the two, these facts are more clearly brought out in graphic

same for the interval 40-49, we have $7 + 11 + 11 + 4 = 33$. Divided by 4, this becomes 8.25. For the interval 30-39, we have

$$11 + 4 + 4 + 0,$$

all divided by 4, which gives us 4.75. If we wish to do so, we may even estimate frequencies in the end classes given, for example, in the interval 20-29. Here we have $4 + 0 + 0 + 0 = 4$, and divided by 4 the outcome is 1.00. All the expected frequencies for this distribution are given in column (3) of Table 5. Their sum is equal to 51, which is a rough check upon the accuracy of computation.

TABLE 5—ORIGINAL AND SMOOTHED FREQUENCIES FOR A DISTRIBUTION OF SCORES IN A SCHOLASTIC-APTITUDE TEST

(1) Scores	(2) f_o	(3) f_e
120-129	0	0 25
110-119	1	0 50
100-109	0	1 00
90- 99	3	2.75
80- 89	5	4 75
70- 79	6	7 75
60- 69	14	10 25
50- 59	7	9 75
40- 49	11	8 25
30- 39	4	4.75
20- 29	0	1 00
Sums	51	51 00

Plotting a Smoothed Distribution.—The final step is to plot the smoothed curve, which we have in Fig. 4. First the obtained frequencies are plotted as circlets in their proper places. It is always well to show these even though we do not draw the curve through them as before. The expected frequencies are next plotted as points. We can probably see by inspection that the smoothing could be improved upon. In drawing the smoothed curve, we do not feel compelled necessarily to touch all the dots. Being concerned with the general shape freed from probably accidental fluctuations, we take the liberty of further smoothing by inspection and by free-hand drawing. If there were too many irregularities, even in the smoothed points we could, of course, repeat the averaging process, but this is usually not wise,

when the number of cases is not too small, a good form is to let a period stand for one individual, a colon stand for two, and an \times stand for five, as in Table 6. When the frequencies are small numbers, the same plan gives an adequate picture if we let an \times or some other letter stand for each individual.

Exercises

1. For each one of the following ranges of measurements, state your judgment of (1) the best size of class interval, (2) the score limits of the lowest class interval, (3) the exact limits of the same interval, and (4) its midpoint

a 83 to 197

b 4 to 39

c 17 to 32

d 35 to 96

e 0 to 188

f -24 to +28

g 0.141 to 0.205

2. Given the following list of scores in a "nervousness" test (Data *A*) and using a class interval of 5, set up a frequency distribution. In the first solution, begin the lowest class interval with a score of 35. List all exact limits of class intervals and also exact midpoints. In a second solution, start the lowest class interval with a score of 33. After finishing both solutions, write out a comparison of the two distributions and defend the choice of the one as against the other. As a third solution, use an interval of 3, choosing your own starting places for the classes. Discuss the relative merits of the third distribution as compared with the first two.

DATA *A*—SCORES IN A NERVOUSNESS INVENTORY

59	48	53	47	57	64	62	62	65	57	57	81	83
48	65	76	53	61	60	37	51	51	63	81	60	77
71	57	82	66	54	47	61	76	50	57	58	52	57
40	53	66	71	61	61	55	73	50	70	59	50	59
69	67	66	47	56	60	43	54	47	81	76	69	

3. Given the following list of scores, each of which is the percentage of 400 words judged pleasant by an individual (Data *B*), set up a frequency distribution making the wisest choice of class interval and class limits.

DATA *B*—AFFECTIVITY RATIOS
(All Have Been Rounded to the Nearest Whole Number)

43	62	52	48	46	65	43	48	52	51	57	48	48
38	42	44	46	43	35	42	42	45	44	46	40	40
47	52	38	51	45	38	51	40	46	45	54	55	41
50	59	42	39	56	44	43	47	51	43	50	34	40
53	42	31	44	51	43	48	41	43	48	41	55	

4. Plot a frequency polygon and a histogram for Data *C*, Group I. State your conclusions about these data as revealed by your plotted distributions.

CHAPTER III

MEASURES OF CENTRAL TENDENCY

This chapter is about averages, of which there are several kinds. Three of them—the *arithmetic mean* (or *mean*, for short), the *median*, and the *mode*—will be explained here. Two others, the *geometric mean* and the *harmonic mean*, are rarely useful to students of psychology and education and will not be discussed.

An *average* is a number indicating the central tendency of a group of observations or of individuals. To the question, “How good is a sixth-grade class in arithmetic?” the most reliable and meaningful kind of answer would be the mean or median in some recognized test of arithmetical achievement. To the question, “What is the weakest tone to which this dog will respond?” the best kind of answer is to state the average result from a number of trials. In either case a single score or a single measurement of the threshold stimulus would be highly unreliable, for not all measurements have the same value. And to answer those questions by reciting the long list of individual measurements would be highly uneconomical in the reporting and not very enlightening to the questioner.

The average, whether it be mean, median, or mode, is therefore first of all a generalization or shorthand description of a mass of quantitative information. It is surely more meaningful and economical to let one number stand for a group than to try to remember all particulars. We also frequently think of a certain kind of group—for example, the sixth-grade group—as having a *true* amount of ability, as of arithmetical achievement that is characteristic of sixth-graders. Or we think of a dog having a *true* or characteristic power to hear minimal sounds. We also believe that the average we obtain from a sample of observations is somewhere near the true mean or that we are estimating what the true central tendency or most characteristic value is for the thing we are measuring when we obtain an average of a sample. The *true* average we can never know; we can only estimate it. But it will be seen later that we can decide about how far any obtained average is from the true average (see Ch. VIII).

X = midpoint of a class interval

f = number of cases within an interval.

The solution by way of this formula is illustrated in Table 7. Here we have only as many X values as there are class intervals instead of as many as there are original measurements. Each class interval has as its X value the midpoint of that interval. This assumes that the midpoint of the interval correctly represents all the scores within that interval. This will not be exactly true in many instances, but the discrepancy is small in any case, and for the purposes of finding the mean, some of the discrepancies counterbalance others, so that the final result is not seriously in error.

TABLE 7—COMPUTATION OF THE MEAN IN GROUPED DATA

(1)	(2)	(3)	(4)
Scores	X Midpoint	f	fX
55-59	57	1	57
50-54	52	1	52
45-49	47	3	141
40-44	42	4	168
35-39	37	6	222
30-34	32	7	224
25-29	27	12	324
20-24	22	6	132
15-19	17	8	136
10-14	12	2	24
Sums		50 N	1,480 ΣfX

$$\text{Mean} = \frac{\Sigma fX}{N} = \frac{1,480}{50} = 29.60$$

In column (2) of Table 7, the midpoints of the intervals are given. We must add each midpoint into our total as many times as there are cases within the interval. This means finding for each interval the product f times X , or fX . The fX products are listed in column (4). The sum of the fX products (ΣfX) is equal to 1,480. Dividing this by N , we find the mean to be 29.60, as it was for the same data ungrouped. As was indicated before, we may expect a minor discrepancy between the means calculated from grouped and ungrouped data. It just happened here that the discrepancy was zero. We may also expect trivial

Choosing a Guessed Mean—First we select a guessed mean. This may be chosen anywhere, for its choice is arbitrary with us. In order to obtain the greatest benefits from the short method, however, it is well to choose a guessed mean rather near to the actual mean, at any rate, somewhere near the center of the distribution. Several criteria guide us in making this choice. One is to place the guessed mean at the midpoint of the middle-class interval (if there is an even number of intervals, either of the two middle ones is eligible). The distribution in Table 8 is distinctly skewed, however, with the bulk of the cases at the lower part of the range; so the mean we find will probably fall in an interval lower than the middle one. Another criterion is to choose the interval containing the median (see page 34 for the method of finding the median). In this distribution, the median falls within the interval for scores 80–89. This is farther from the center than we would ordinarily go for the guessed mean. Another guide is to choose an interval that has a large number of cases—in fact, the largest number. Here it is the interval for scores 90–99. As a good compromise among all these criteria, the interval labeled 90–99 seems best. We should actually come out with the same computed mean no matter which interval we chose for the guessed mean; the choice is dictated entirely by the desire to keep the numbers small so that “headwork” can replace paper-and-pencil work as much as possible.

The Size of Class Interval Becomes the Temporary Working Unit.—Having chosen the interval 90–99, we guess the mean to be at the midpoint of this group, the midpoint being 94.5 (midway between 89.5 and 99.5). The score point of 94.5 becomes the temporary zero point for our measuring scale. In column (3), a zero is written in line with the interval whose midpoint is 94.5. The first interval above is given a value of +1; the second, +2; the third, +3; etc. The first interval below is given a value of -1; the second, -2; etc. These x' values now represent the class intervals, which are just one unit apart. The new unit is equivalent to 10 score-point units, a fact that we shall have to remember later.

The Correction in the Guessed Mean.—From here on, the next steps are similar to those taken in Table 7. Next we find the fx' product for each interval, *taking great care to record algebraic signs*. All products above the guessed mean are positive, and all products below are negative. The sum of the positive products is +56, and the sum of the negative products is -68. The algebraic sum of the entire column is therefore $56 - 68$, which equals -12. The $\Sigma fx'$ therefore equals -12. From this we can find directly how far the actual mean is from

our guessed mean. The actual mean is equal to M' plus a correction c , and this correction is given by the formula

$$c = i \left(\frac{\sum fx'}{N} \right) \quad (3)$$

where i = size of the class interval.

x' = deviation of a class interval from the guessed mean in terms of i as the unit.

f = frequency within a class interval.

N = total number of measurements.

In this problem, $i = 10$, $\sum fx' = -12$, and $N = 64$. Therefore

$$c = 10 \left(\frac{-12}{64} \right) = -\frac{120}{64} = -1.88$$

Adding this correction to the guessed mean, we have

$$M' + c = 94.5 - 1.88 = 92.62$$

The mean is 92.62 score points, but we should usually report it merely as 92.6 score points.

A Summary of the Short Solution of the Mean.—The steps involved in the short method of computing the mean may be summarized as follows:

- Step 1. Set up the frequency distribution.
- Step 2. Choose a guessed mean. This is the midpoint of the interval (1) near the center of the distribution; or (2) containing the median or mode or both; and (3) probably containing the actual mean.
- Step 3. Assign to the class intervals new small integral values, starting with zero at the interval containing the guessed mean, with positive values above and negative values below. Call these new values x' .
- Step 4. Find the fx' product for each interval, and record in a column.
- Step 5. Sum the fx' products algebraically. This is $\sum fx'$.
- Step 6. Divide the sum of the fx' products by N .
- Step 7. Multiply this quotient by i , the size of the class interval. This is the correction c .
- Step 8. Add this correction algebraically to the guessed mean. This gives the mean.

THE MEDIAN

The *median* is defined as that point on the scale of measurement above which are exactly half the cases and below which are the other half. Note that it is defined as a *point* and not as a score or any particular measurement. If this conception is kept clearly in mind, many difficulties will be forestalled. Some textbooks on statistics give a different definition of median for ungrouped as compared with grouped data and recommend two different procedures for computing the median. Here we shall apply the same definition to both cases and be consistent in computation throughout.

The Median from Grouped Data.—It is probably easier to grasp the process of computing a median in grouped data. For a first illustration, consider Table 9. Here there are 28 cases; so the median

TABLE 9—COMPUTATION OF THE MEDIAN SIZE OF CLASS IN A CERTAIN SCHOOL, WITH THE USE OF GROUPED DATA

Class size	<i>f</i>	
40-44	1	
35-39	0	
30-34	3	
25-29	5	
20-24	3	12 = number of cases above the interval containing the median
15-19	10	
10-14	1	
5- 9	1	6 = number of cases below the interval containing the median
0- 4	4	
<i>N</i> = 28		

$$Mdn = 14.5 + \frac{8}{10} \times 5 = 14.5 + 4.0 = 18.5$$

$$Mdn = 19.5 - \frac{2}{10} \times 5 = 19.5 - 1.0 = 18.5$$

is that number of points on the measuring scale above which there are 14 and below which there are 14. Counting frequencies from the bottom upward, we find that $4 + 1 + 1 + 10 = 16$ cases, or 2 more than we want. To make 14 cases, we need 8 out of the 10. The median lies somewhere within the interval 15-19, whose *exact* limits are 14.5 and 19.5. We assume for the sake of computation that the 10 cases within this interval are evenly spread over the distance from 14.5 to 19.5 (see Fig. 5). We must interpolate within this range to find

how far above 14.5 we need to go in order to include the 8 cases we need below the median. We must go $8/10$ of the way, for 8 is the number we require, and 10 is the total number in the interval. The total distance is 5 units; so on the scale of measurement we go $8/10$ of 5, or exactly 4.0 units. Adding this 4.0 to the lower limit of the class interval 14.5, we get

$$14.5 + 4.0 = 18.5$$

as the median.

We can always check this by counting down from the top of the distribution until we include $N/2$ of the cases; 14 in this problem. Starting at the top, we find that

$$1 + 0 + 3 + 5 + 3 = 12$$

We need 2 more cases out of the next group of 10. We must go $2/10$ of the way below the *upper* limit of the interval, that is, 19.5. This means $2/10$ of 5 or exactly 1.0 unit. The upper limit, 19.5 minus 1.0 gives us 18.5 for the median, which checks with the

one obtained by counting up from below. It is well always to check the determination of a median in this manner, and to do so involves very little work. If the two estimates do not come out exactly the same, something is wrong.

To take another example with grouped data, consider Table 10, where N is an odd number. Here $N/2$ is 18.5, but the principle of interpolating within an interval for the exact median is just the same. Counting up from below, we find that $1 + 5 + 8 = 14$, which lacks 4.5 cases of including the lower half. In the next interval, we must go $4.5/8$ of the way, or $4.5/8$ times 2, which equals $9/8$, or 1.125. Adding this many units to the lower limit of the interval (22.5), we have 23.625 as the median; or dropping all but one decimal place, we report the median as 23.6 score points. Checking by counting down from the top, we find 15 cases above the point 24.5. Going $3.5/8$ of the way down into the interval of 2 units, we find that we must deduct 0.875

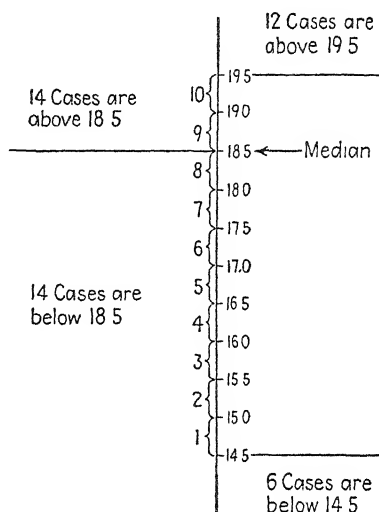


FIG 5—Showing how the 10 cases in the interval 14.5 to 19.5 are distributed, each assumed to occupy a tenth of the interval or one-half of a score unit. The eighth one extends up to the point 18.5, which is the median.

TABLE 10—COMPUTATION OF THE MEDIAN SCORE IN A SENTENCE-CONSTRUCTION TEST AS GIVEN TO 37 MEN

Scores	<i>f</i>	
37-38	1	
35-36	2	
33-34	0	
31-32	1	
29-30	0	
27-28	6	15 = number of cases above interval containing the median
25-26	5	
23-24	8	
21-22	8	14 = number of cases below interval containing the median
19-20	5	
17-18	1	
$N = 37$		$N/2 = 18.5$

$$Mdn = 22.5 + \frac{4.5}{8} \times 2 = 22.5 + \frac{9}{8} = 22.5 + 1.125 = 23.6$$

$$Mdn = 24.5 - \frac{3.5}{8} \times 2 = 24.5 - \frac{7}{8} = 24.5 - .875 = 23.6$$

from 24.5 to find the median. When rounded to one decimal place, the median is 23.6, as before. In terms of a formula, the interpolated median is found from below by

$$Mdn = l + \left(\frac{\frac{N}{2} - F_e}{f_p} \right) i \quad (4)$$

where l = exact lower limit of the class interval containing the median.

F_e = sum of all frequencies below this point.

f_p = frequency of the interval containing the Mdn .

N and i are defined as usual.

A Summary of the Steps for Interpolating a Median.—The steps for computing a median from grouped data may be summarized as follows:

- Step 1. Find $N/2$, or half the number of cases in the distribution.
- Step 2. Count up from below until the interval containing the median is located.
- Step 3. Determine how many cases are needed out of this interval to make $N/2$ cases.

- Step 4. Divide this number needed by the number of cases within the interval.
- Step 5. Multiply this by the size of class interval.
- Step 6. Add this to the exact lower limit of the interval containing the median.
- Step 7. Check by adding down from the top to find to what point the upper half of the cases extend in a manner analogous to that described in Steps 2 to 5 inclusive.
- Step 8. Deduct the number of score units found in Step 7 from the exact upper limit of the interval containing the median.

Some Special Situations.—There are some instances in which things do not turn out just as they did in the two illustrative examples.

When the Median Falls between Intervals.—If it should happen, in adding up cases from below, that half the cases take in *all* the cases in the last interval, the median is then the exact upper limit of that interval. In counting down from above, it would be found that all the cases in the interval just above this one would also be required to make $N/2$; so its exact bottom limit would be the median. This coincides with the exact upper limit of the interval below; so the median checks. As an example, note the following fictitious data:

Scores	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59
<i>f</i>	2	7	10	15	18	8	3	5

Here $N/2$ is 34. This many cases takes us exactly through the interval 35-39. The median is 39.5. From above down, we are carried through the interval 40-44, whose lower limit is 39.5. Again the median is 39.5.

When There Are No Cases within the Interval Containing the Median. Another question arises when the median falls within an interval where there are *no* cases. It is even possible that in the region of the median, two or more intervals have frequencies of zero. In this case, the median is taken as *the middle of the range having no cases*. If that range is one interval, the median is taken as the midpoint of that interval; if it covers two intervals, the median would be the division point between those intervals, etc.

Scores	5-7	8-10	11-13	14-16	17-19	20-22	23-25	26-28
<i>f</i>	1	7	9	0	6	7	2	2

In the data just preceding, the median is 15.0, which is midway between 13.5 (to which point the lower half of the cases extend) and 16.5 (to which point the upper half of the cases extend). Or it is the arithmetic mean of those two limits, for $16.5 + 13.5$ divided by 2 is 15.0.

The Median from Ungrouped Data.—Things learned in finding a median in grouped distributions should carry over almost intact to the use of ungrouped data. The median is a *point* on the measuring scale. In ungrouped data, each score or measurement is assumed to occupy a range of one unit. The median either falls within one of those units or somewhere between units. The first step is to arrange the measurements in order of their size. The list of 10 measurements of the threshold for pitch as given on page 29, when placed in rank order, becomes

11, 11, 11, 11, 13, 13, 13, 15, 17, 17

As in the case of grouped data, it is assumed that the four 11's occupy the range from 10.5 to 11.5, the three 13's occupy the range from 12.5 to 13.5, etc. Counting from below to include 5 cases brings us to the first 13 that must be included among the 5. We must therefore extend $1/3$ of the way in the interval of 1 unit, or 0.33 unit into the interval, starting at 12.5. The median is $12.5 + 0.33$, which equals 12.83, or, when rounded, 12.8. In checking from above, the median is found at $13.5 - 0.7$, which also equals 12.8.

In the series of measurements

2, 5, 7, 8, 9, 10, 17

the median comes midway in the fourth one, which is 8. Since 8 occupies a range of 7.5 to 8.5, the median is the midpoint of this range, or exactly 8.0. In the series of measurements

7, 9, 10, 12, 13, 15, 18, 20

four are 13 or above, and four are 12 or below. The division between upper and lower halves comes as 12.5, which is the median in this case. In the array of scores

15, 17, 18, 20, 23, 24, 27, 30

the lower half extends up to 20.5, and the upper half extends down to 22.5. Midway between these two values is the point 21.5, or the average of the two.

It is probably obvious that the median of so small a number of observations cannot be very reliable, and we should not place too much reliance upon it or carry our calculations to more than one decimal

place (we might even report nearest whole numbers); but in order to keep consistent certain principles of the median and of the process of computing it, certain steps have been emphasized. Whenever there is doubt concerning special cases not covered in these illustrations, an application of these principles should take care of the matter.

THE MODE

The *mode* is strictly defined as *the point of maximum frequency in a distribution*. When we have ungrouped data, the mode is that measurement which occurs most frequently. Usually it is somewhere near the center of the distribution, and in a strictly normal distribution it coincides with the mean and the median.

The Crude Mode.—*In a distribution of grouped data, the crude mode is the midpoint of that class interval having the greatest frequency.* In Table 7, the highest frequency is 12, for the interval 25–29. The midpoint of this interval is 27; so the mode is taken to be 27.0. In Table 8, there are two intervals with the same maximum frequency of 12. If these two intervals had been separated by more than one intervening interval of lower frequency, we should be justified in saying that the distribution is *bimodal* (having two modes). But the single intervening frequency of 10 hardly gives us sufficient basis for this conclusion. The distribution is therefore probably really unimodal, but we are not able to decide upon its crude mode. A calculated mode can be found, as we shall soon see.

In Table 9, the crude mode is clearly 17.0. In Table 10, the maximum frequency is shared by two neighboring intervals. In a situation like this, we do the reasonable thing of assigning the crude mode to the dividing point between these intervals, which is 22.5. Unless the data are reasonably numerous, so that there is clearly an interval of highest frequency, we should not attempt to assign a modal value to the distribution. For example, the 10 measurements of threshold for pitch present an unusual situation with the greatest frequency (four cases) of 11, which is at one end of the distribution. Following right behind is the measurement 13, with three cases. Here it would be rather meaningless to say that the mode is 11.

The Mode Estimated by Computation.—Fortunately, because of certain mathematical relationships between the mode and the other two measures of central tendency, we can estimate it from them. A simple approximate formula is

$$Mo = 3Md_n - 2M \quad (5)$$

In other words, the mode equals three times the median minus two times the mean.

Applying this formula, we can now estimate the mode of the distribution in Table 8, in which we were unable to decide upon a crude mode. The median for this distribution is 88.5, and the mean is 92.62. Although we rounded the mean to one decimal place in reporting it, in further calculations with it, we do well to keep the second decimal place. Applying formula (5)

$$(3 \times 88.5) - (2 \times 92.62) = 265.5 - 185.24 = 80.26$$

Rounded to one decimal place, the estimated mode is 80.3. Reference to the distribution in Table 8 again will show that this point comes about midway among the four high frequencies. Had we done a very reasonable thing and placed the crude mode midway among these four intervals, it would have been at 79.5, which is less than one unit from the calculated mode.

The mean of the distribution in Table 9 is 19.14 and the median is 18.5. The calculated mode is $(3 \times 18.5) - (2 \times 19.14)$, which equals $55.5 - 38.28$, or 17.22. This is separated from the crude mode, which is 17.0, by a trivial amount. In the distribution in Table 10, the median is 23.6, and the mean is 24.52. From this information, the mode is estimated as 21.8, which deviates from the crude mode only 0.7 unit. It may add meaning to the computed mode to say that it is the point on the measuring scale at which the smoothed distribution curve probably has its highest point.

WHEN TO EMPLOY THE MEAN, MEDIAN, AND MODE

Certain Advantages of the Mean.—The arithmetic mean is to be preferred whenever possible because of several desirable properties. In the first place, it is generally the most reliable or accurate of the three measures of central tendency. By this we mean that from sample to sample of the same population, the mean will ordinarily fluctuate less widely. Another reason is that the mean is better suited to further arithmetical computations. Deviations of single cases from the central tendency are important information about any distribution. Much is done with these deviations, as will be seen in the following chapter. It will also be found that we square those deviations, and this we are really justified in doing only when the deviations are taken from the mean. When distributions are reasonably symmetrical, we may almost always use the mean and should prefer it to the median and mode. On the other hand, there are instances, particularly when

distributions are skewed and when the mean would lead to erroneous ideas about a distribution, in which other measures of central tendency are better used.

A Comparison of the Mean with Median and Mode.—One property of the mean is that it is sensitive to the size of extreme measurements when they are not balanced by other extreme measurements on the other side of the middle. In the following set of measurements, the mean is 9 and the median is 9:

4, 5, 7, 9, 11, 13, 14

Now, if the 14 had been 23 instead of 14, the median would be unchanged, but the mean would become 10. There are still an equal number of cases above and below 9. So far as the median is concerned,

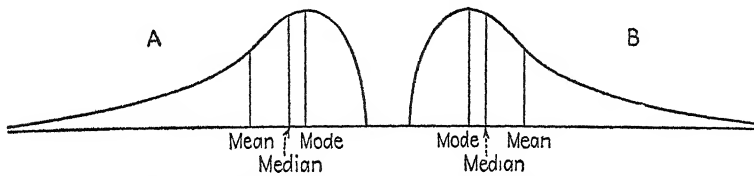


FIG. 6.—Two skewed distributions, (A) skewed negatively, and (B) skewed positively, showing the relative positions of modes, medians, and means

the 11, 13, and 14 could have been 110, 130, and 140, and still the median would be 9. But in this rather unusual but not impossible event, the mean would become 57.9, where formerly it was only 9. The conclusion to be drawn is that when there are any very extreme measurements not balanced by other extreme measurements in the other direction, the median is to be preferred to the mean.

Central Tendencies in Skewed Distributions.—In general terms, we are dealing here with skewed distributions. In these, the mean is always pulled toward the skewed (pointed) end of the curve, as Fig. 6 shows. The arithmetic mean comes at the center of gravity of the distribution. The *sum* of the scores on the one side of it equals the *sum* of the scores on the other side. The median comes at a point that divides the area under the distribution curve into two equal parts. The *number* of scores on the one side of it equals the *number* of scores on the other. The interpretations of mean and median should be made accordingly. For example, for the data on class size in Table 9, the *median* of 18.5 tells us that half of the classes had 19 or more students enrolled and half of them had 18 or less. The mean class size, which is 19.14, tells us that if all the enrolled students had been reapportioned

so as to make all classes the same size, the enrollment in each class would have been 19.14, or 19, with a few students left over.

When the Mean Is Misleading.—In some instances, to give the mean of a distribution only is highly misleading, for example, in a study of class size in a certain university, among 62 classes, there were 2 classes having more than 200 students, and 2 having between 100 and 200 students, all the remaining classes except 2 being smaller than 60. The average size of the 62 classes was 34, but this was not very typical, because half of the classes had 20 or less (the median was 20.5). The most *typical* size of class would be given as the *mode*, which was 17 (crude mode). If our purpose happened to be to equalize the size of classes, assuming that this were practical, we could conclude that there would be 34 students per class. If we wanted to decide as a matter of educational policy whether or not there were too many small classes in general and if we had concluded beforehand that most teachers can successfully handle 30 students in a group, then the median would tell us, without knowing anything more about the distribution, that there were entirely too many small classes. The mean would not have told us this, because it was higher than 30. If we were piloting a visiting inspector about the buildings while classes were in session and wished to prepare him for the most likely size of class he would find at random, we should give him the mode, since this size is more likely to occur than any other one size. If we were purchasing equipment to suit classes of various sizes, we should adapt it, if necessary, most often to classes of modal size, though in this case we should also want to know more about the entire frequency distribution.

Mean and Median Often Both Reported.—In reporting upon central tendencies of skewed distributions, it is usually well to state both the mean and the median, since each tells its own story, and from the difference between the two we can immediately infer in what direction the distribution is skewed and about how strongly. Although the mode is easily and quickly determined and will often serve until better averages can be computed, it should probably never be reported alone and need not be reported with the other two averages except when it is meaningful to do so. When a distribution is symmetrical about the mode, the three averages will coincide, and so only one of them, preferably the mean, need be reported, together with the fact that the distribution is symmetrical.

When the Median Is Especially Called For.—There are one or two kinds of distribution in which the median is the only satisfactory average.

When Distributions Are Truncated.—One of these kinds is the *truncated* distribution. This is the type in which the exact values of extreme cases are not known. In certain work-limit tests, for example, some subjects would work on for unusual lengths of time if permitted to do so. Suppose that all those who work on a certain test up to 10 min. are arbitrarily stopped. They are in the minority; so a median can be found. Time spans up to 10 min. may be classified as usual into chosen class intervals. From 10 min. up, we find the laggards grouped together. We do not know just how long they might have kept working had we let them continue. An arithmetic mean cannot be determined here, but median and mode can still be utilized.

When Equality of Unit Is Uncertain.—In another instance, we are not sure that all the units of our measuring scale are equal. This is particularly true in the psychological scaling methods of rank order and of equal-appearing intervals. In the former case, a number of judges have placed several objects or persons in rank order for some quality. Though the ranks are numerically equidistant, the things ranked probably are not. When combining ranks for any one object, we do less violence to the measurement if we find a median rather than a mean. In the other instance, though objects are placed in piles or categories that seem equidistant to the observer, again we are not sure that his categories are numerically equidistant, and the median is a safer statistic to compute. It is also true in this scaling method that distributions of judgments for objects very high or very low on the scale are skewed or even truncated because of the "end effect." By the end effect, we mean that although some judges would like to place some stimuli above the highest pile or category (or below the lowest), they are not permitted to do so. Some objects or persons rated thus pile up in the end categories when some of these times they should have gone beyond the end. This fact will distort the arithmetic mean but will not influence the median so long as not more than half of all the judgments for an object fall in the end group.

A Summary of When to Use the Three Averages.—In brief, the following rules will generally apply:

1. *Compute the arithmetic mean when*
 - a. The greatest reliability is wanted.
 - b. Other computations, as finding measures of variability, are to follow.
 - c. The distribution is symmetrical about the center, particularly when it is approximately normal.

2. *Compute the median when*

- a. There is not sufficient time to compute a mean.
- b. Distributions are badly skewed. This is the case when one or more extreme measurements are at one side of the distribution.
- c. We are interested in whether cases fall within the upper or lower halves of the distribution and not particularly in how far from the central point.
- d. An incomplete (truncated) distribution is given.
- e. There is uncertainty about the equality of the unit of measurement.

3. *Compute the mode when*

- a. The quickest estimate of central tendency is wanted.
- b. A rough estimate of central tendency will do.
- c. We wish to know what is the most typical case.

Exercises

DATA D.—SCORES IN AN ENGLISH-USAGE

EXAMINATION	
Scores	<i>f</i>
52-53	1
50-51	0
48-49	5
46-47	10
44-45	9
42-43	14
40-41	7
38-39	8
36-37	6
34-35	5
32-33	3
Sum.	<u>68</u>

DATA E—AFFECTIVITY SCORES
(Per Cent of 400 Words Marked
"pleasant" or "unpleasant")

Scores	<i>f</i>
95-99	6
90-94	11
85-89	16
80-84	7
75-79	9
70-74	8
65-69	2
60-64	3
55-59	2
50-54	1
Sum.	<u>65</u>

DATA *F*—AGES OF COLLEGE FRESHMEN

Age at last birthday	Men	Women
31-35	1	2
26-30	3	6
25	7	6
24	6	7
23	11	7
22	20	6
21	23	16
20	40	13
19	88	48
18	117	67
17	69	57
16	2	6
Sums	387	241

DATA *G*—AIMING-TEST SCORES
(In Terms of Average Error
in Millimeters)

Score	Men	Women
8 0-8 4	1	
7 5-7 9	5	
7 0-7 4	2	
6 5-6 9	7	2
6 0-6 4	6	4
5 5-5 9	11	3
5 0-5 4	10	9
4 5-4 9	16	7
4 0-4 4	18	15
3 5-3 9	19	12
3 0-3 4	17	15
2 5-2 9	17	13
2 0-2 4	14	14
1 5-1 9	13	10
1 0-1 4	8	1
0 5-0 9	1	
Sums .	165	105

1 Compute the arithmetic mean of any or all distributions in Data *D* to *G* inclusive, using the method that seems most feasible. In Data *F*, you will need to make some assumption about the cases in the two highest intervals. State your assumptions if means are computed for these distributions.

2 Compute medians for any or all distributions in Data *D* to *G* inclusive. Why is the difficulty experienced with computation of the mean in Data *F* not also encountered in computing the median?

3. Give the crude modes for all distributions in Data *D* to *G*. Compute the estimated mode in distributions for which you know both mean and median.

DATA *H*—SOME UNGROUPED DATA

a 8, 15, 13, 6, 10, 16, 7, 12, 11, 14, 9

b 12, 10, 18, 13, 4, 8, 17, 15, 6, 14

c 9, 8, 9, 15, 3, 9, 11, 9, 13

d 12, 28, 19, 15, 15, 35, 14, 15

e 7, 18, 20, 14, 27, 23, 13, 3

4. Compute and list the means, medians, and crude modes (where possible) for the distributions in Data *H*.

5. For each distribution in Data *H*, tell to which measure of central tendency you give first preference and to which, second. Give reasons.

6. For each distribution in Data *C* to *G* inclusive, tell which measure of central tendency you would prefer and which would be your second choice. Give reasons.

CHAPTER IV

MEASURES OF VARIABILITY

Knowing the central tendency of a set of measurements tells us much, but it does not by any means give us the total picture of the sample we have measured. Two groups of six-year-old children may have the same average IQ of 105, from which we would conclude that, taken as a whole, each group is as bright as the other, and we might expect from the two the same average level of performance in school or out of school in areas of life where IQ is important. Yet when we are told, in addition, that one group has no individuals with IQ 's below 95 or above 115, whereas the other has individuals with IQ 's ranging from 75 to 135, we recognize immediately that there is a decided difference between the two groups in variability or dispersion of brightness. The first group is decidedly more homogeneous with respect to IQ , and the second is decidedly more heterogeneous. We should expect the first group to be much more teachable in that they will grasp new ideas at about the same rate and progress at about the same rate. We should expect the second group to show considerable disparity in speed of grasping new ideas. There will be extreme laggards at the one end of the distribution and others at the other end of the distribution who may be irked at the slow progress of the group. The distributions for two such groups, when plotted, resemble those in Fig. 7.

It is the purpose of this chapter to explain and illustrate the methods of indicating degree of variability or dispersion by the use of single numbers, just as in the preceding chapter we saw how the central tendency of a distribution could be indicated by a single number. The five customary values to indicate variability are (1) the total range, (2) the middle 80 per cent range, (3) the semi-interquartile range Q , (4) the standard deviation σ , and (5) the average (or mean) deviation AD .¹

¹ The *probable error* PE has traditionally been used as a measure of variability, but it seems rapidly to be going out of use and so is merely mentioned in this volume (see the footnote on p. 55)

THE TOTAL RANGE

The total range is the indicator of variability that is easiest and most quickly ascertained but is also the most unreliable, and so it is almost entirely limited to the purpose of preliminary inspection. In the illustration of the preceding paragraph, the range of the first group (from an *IQ* of 95 to an *IQ* of 115) was 21 *IQ* points inclusive. The range of the second group was from 75 to 135 *IQ* points. The range is the distance given by highest score minus lowest score, plus 1. From this comparison, we draw the conclusion that the second group is about three times as variable as the first.

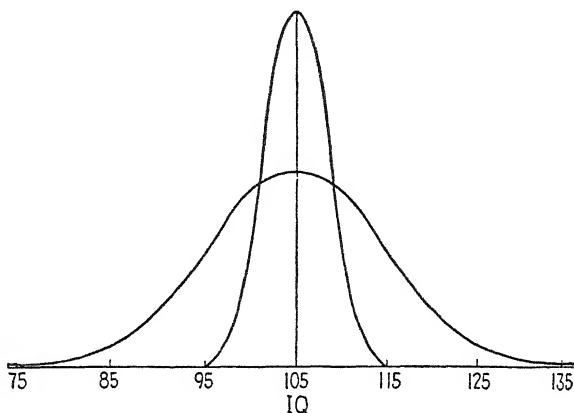


FIG. 7.—Two distributions with the same mean (*IQ* = 105) but with decidedly different ranges.

Why the Range Is Unreliable.—But the range is often very unreliable for the reason that only two measurements alone determine it. The remaining measurements would then have nothing whatever to do with the matter. In the second group just mentioned, it might have been true that there were several *IQ*'s of 75 and also several *IQ*'s of 135; but this would be most unusual. The chances are great that there would be only one 75 and one 135. Furthermore, the next lowest *IQ* might have been 85, with a gap of 10 points to the very lowest; and the next to the highest might have been 120, a distance of 15 points from the very highest. Had either or both of the persons with 75 *IQ* and 135 *IQ* been missing from the group, the range would have been something very different from the 61 points actually obtained. This is what we mean by saying that the total range is highly unreliable. Some faith can, of course, be placed in it when there is more

than one case having each of the extreme measurements and when there are no decided gaps in the tails of the distribution.

When Ranges Should Not Be Compared.—Total ranges should not be compared when two distributions have a markedly different number of cases. It is quite natural for more extreme cases to show up as we add new cases to any sample, so that larger groups should be expected to have wider total scatter. This factor is not nearly so important for other indicators of dispersion as it is for total range. Another caution almost goes without saying, and that is the impossibility of comparing ranges in two distributions where the units of measurement are not the same.

THE MIDDLE 80 PER CENT RANGE

A much more stable idea of range is given by finding how much of the measuring scale lies between the two points between which are the middle 80 per cent of the cases. Obviously, 10 per cent of the cases lie below this range, and 10 per cent lie above it. The points are determined in much the same manner as the median was found in the preceding chapter—by interpolation. In finding the median, we counted up (or down) to include exactly 50 per cent of the cases. In this problem, we count up from the bottom (and down from the top) to include the lowest and highest 10 per cent of the cases.

For our example, let us return to the ink-blot-test scores that are given in Table 11. Since there are 50 cases in the distribution, 10 per

TABLE 11.—DETERMINATION OF THE MIDDLE 80-PER CENT RANGE FOR THE DISTRIBUTION OF INK-BLOT SCORES

Scores	<i>f</i>
55-59	1
50-54	1
45-49	3
40-44	4 ← Highest 10% above this point
35-39	6
30-34	7
25-29	12
20-24	6
15-19	8 ← Lowest 10% below a point within this interval
10-14	2

Top of the 80 per cent range falls at 44.5.

Bottom of the 80 per cent range is interpolated as follows:

$$14.5 + \frac{3}{8} \times 5 = 14.5 + 1.875 = 16.375$$

Middle 80 per cent range is $44.5 - 16.375 = 28.125$, or 28.1

cent is exactly 5. Counting up from the bottom to include 5 cases, we find that we must go $3/8$ of the way into the second class interval. This means $3/8$ of 5, which is 1.875. Added to the exact lower limit of the second interval, which is 14.5, we have 16.375. At the upper end of the distribution, we find that 5 cases carries us down exactly between the third and fourth interval. The division point between these two intervals is 44.5. The range we are seeking is 44.5 minus 16.375, which is 28.125, or, rounded to one decimal place, it is 28.1. This distance on the measuring scale can now be compared with the 80 per cent range of any other group *on the same measuring scale*. If, for example, a second group gave us an 80 per cent range of 35.2, we should be able to conclude that it is about 25 per cent more variable than the first, for its range is about 7 points greater, and 7 is 25 per cent of 28.

THE SEMI-INTERQUARTILE RANGE Q

The middle 80 per cent range, though more reliable than the total range, is still not nearly so reliable as other indicators of variability. Its terminal points are determined among the tails of the distribution where cases are relatively sparse. Other measures of variability are determined closer to the central mass of cases, or even by the entire sample, and so are more reliable. One of these, as readily determined as the 80 per cent range, is the semi-interquartile range Q . This is *one-half* the range of the middle 50 per cent of the cases. First we find by interpolation the range of the middle 50 per cent or interquartile range, then divide this range by 2.

Quartiles and Quarters.—When we count up from below to include the lowest or first quarter of the cases, we find the point called the *first quartile*, which is given the symbol Q_1 . Counting down from above to include the highest or fourth quarter of the cases, we locate the third quartile, or Q_3 . Incidentally, the median, which separates the second and third quarters of the distribution, is also called Q_2 . Note that the quartiles Q_1 , Q_2 , and Q_3 are *points* on the measuring scale. They are the division points between the *quarters*. We may say of an individual that he is *in* the highest quarter (or fourth quarter), and we may say of another that he is *at* the third quartile. We should never say of an individual that he is in a certain quartile.

Interpolation of Q_1 and Q_3 .—In the distribution of ink-blot scores again, we locate the third and first quartiles by interpolation (see Table 12). One-fourth of the cases ($N/4$) is 12.5. Counting up from the bottom to include 12.5 cases, we find that we need 2.5 out of the 6 cases in the third class interval. As in earlier solutions, $2.5/6$ times

5 gives 2.08. Added to 19.5, this gives 21.58 as the position of Q_1 . Counting down from the top, we find that we need 3.5 cases out of 6 in the fifth class interval. So $3\frac{5}{6}$ of 5 gives 2.92. Deducted from 39.5, this leaves 36.58 as our estimate of Q_3 .

TABLE 12.—DETERMINATION OF Q_3 , Q_1 , AND Q (THE SEMI-INTERQUARTILE RANGE) FOR THE INK-BLOT TEST SCORES

Scores	f
55-59	1
50-54	1
45-49	3
40-44	4
35-39	6 ← Q_3 lies within this interval
30-34	7
25-29	12
20-24	6 ← Q_1 lies within this interval
15-19	8
10-14	2
$N = 50$	

$$Q_1 = 19.5 + \frac{2.5}{6} \times 5 = 19.5 + 2.08 = 21.58$$

$$Q_3 = 39.5 - \frac{3.5}{6} \times 5 = 39.5 - 2.92 = 36.58$$

$$Q = \frac{36.58 - 21.58}{2} = \frac{15.00}{2} = 7.5$$

The Interquartile Range and Q —The interquartile range, or the distance from Q_1 to Q_3 , is given by $Q_3 - Q_1$, or $36.58 - 21.58$, which equals 15.00. The semi-interquartile range is one-half of this, or 7.5. In terms of a formula

$$Q = \frac{Q_3 - Q_1}{2} \quad (6)$$

where Q = semi-interquartile range.

Q_3 = third quartile.

Q_1 = first quartile.

How Quartiles Indicate Skewness.—It is of interest in passing to take note of the relative distances of Q_3 and Q_1 from the median, or Q_2 , in a distribution. If the distribution is exactly symmetrical, both the third and first quartiles will be the same distance from the median, and that distance is Q . When there is any skewness in the distribution, the two distances will be unequal. If the skewness is positive, the distance $Q_3 - Q_2$ will be greater than the distance $Q_2 - Q_1$. If the

skewness is negative, the reverse will be true. The relative sizes of these two distances therefore tells much about the direction and the amount of skewness in the distribution. For the ink-blot scores, $Q_3 - Q_2$ is 8.4, and $Q_2 - Q_1$ is 6.6. Our inference is that the distribution is positively skewed to a moderate degree.

THE STANDARD DEVIATION

The standard deviation, or σ , is the most commonly used indicator of variability, and of the ones described here it is the most reliable. It is also the most difficult to compute; there are several ways in which it may be computed, depending upon the nature of the data and the tools one has available for computation.

The Standard Deviation Computed Directly from Deviations.—In the first method of computation, we begin with the deviations from the mean. Every score or measurement in a distribution deviates from the mean in that it is a certain distance above or below the mean. When and if any measurement coincides exactly with the mean, its deviation is zero. The deviation of any measurement from the mean is given by the simple formula

$$x = X - M \quad (7)$$

where x = deviation from the mean.

X = original score.

M = arithmetic mean.

When X is numerically larger than the mean, x will be positive in algebraic sign; and when X is numerically smaller than the mean, x will have a negative sign. If we were to sum all the deviations from the mean in any distribution, keeping algebraic signs, the sum would be zero. In other words, $\Sigma x = 0$.

Finding a Mean of the Deviations.—In a distribution with great variability, the size of the deviations will run generally large, and in a distribution of low variability, the size of the deviations will run small. We need some kind of average of the deviations to indicate in one summarizing number how large or how small the deviations are. *The standard deviation is a kind of average of the deviations.* But it is not a simple average. Before we find an average of the deviations, we square them. One advantage of this step is that it makes all numbers positive, since a deviation like -4 times itself is $+16$. The simplest formula for the standard deviation reads

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}} \quad (8)$$

where σ = standard deviation

x = any deviation from the mean

Σx^2 = sum of the squared deviations.

N = number of cases, as usual

Summary of the Steps of Computation—The steps necessary to solve for a standard deviation with the use of this formula are

- Step 1. Determine all the deviations by the formula $x = X - M$.
- Step 2. Square every deviation, finding x^2 .
- Step 3. Sum the squared deviations, finding Σx^2 .
- Step 4. Find the mean of the squared deviations, or $\Sigma x^2/N$.
- Step 5. Find the square root of this mean, which is σ , the standard deviation.¹

As an illustrative problem, let us take the 10 measurements of the threshold for pitch (see Table 13). Their mean we found to be 13.2.

TABLE 13—CALCULATION OF THE STANDARD DEVIATION IN UNGROUPED DATA

(1)	(2)	(3)
X Scores	x Deviations	x^2
13	-0 2	04
17	+3 8	14 44
15	+1 8	3 24
11	-2 2	4 84
13	-0 2	04
17	+3 8	14 44
13	-0 2	04
11	-2 2	4 84
11	-2 2	4 84
11	-2 2	4 84
		51 60
		Σx^2

$$\sigma = \sqrt{\frac{51\ 60}{10}} = \sqrt{5.160} = 2.27, \text{ or } 2\ 3$$

The deviations from the mean are given in column (2) and their squares, in column (3). Their sum is 51.60. The mean of the squared deviations is 5.160. The standard deviation is the square root of this,

¹ In this and in all subsequent work, when the square-root sign (or radical) enters into an equation, whatever is under the radical should be reduced to a single number before the square root is extracted.

or 2.27. But since there are only two groups of significant figures (the 5 and the 16) in the number whose square root is to be extracted (see page 12), we should round to one decimal place and report σ as equal to 2.3. In terms of the unit of our measuring scale, this is 2.3 cycles per second.

The Interpretation of a Standard Deviation.—Now that we have the answer 2.3 cycles per second, how shall we interpret it? The usual and most accepted interpretation is in terms of the percentage of cases included within one standard deviation below the mean to one standard deviation above the mean. This range on the scale of measurement includes about two-thirds of the cases in the distribution. In a normal distribution, it is known that from -1σ (one standard deviation below the mean) to $+1\sigma$ (one standard deviation above), exactly 68.26 per cent of the cases are found. Since most samples yield distributions that depart to some degree from normality, we say, "about two-thirds," which is, of course, a trifle short of 68.26 per cent.

In the problem just solved, where we found σ equal to 2.3, the distance from -1σ to $+1\sigma$ on the scale of measurement is 10.9 to 15.5 cycles, *i.e.*, the mean 13.2 minus 2.3 is 10.9, and the mean plus 2.3 is 15.5 cycles. Within these limits are all measurements of 11, 12, 13, 14, and 15. By actual count, there are four 11's, three 13's, and one 15, or 8 of the 10 measurements within these limits, whereas we should have expected 7. But because of the small number of cases and the fact that the distribution is irregular, we should not be surprised at this result.

TABLE 14—CALCULATION OF THE STANDARD DEVIATION IN GROUPED DATA WITH THE USE OF ACTUAL DEVIATIONS

(1)	(2)	(3)	(4)	(5)
X	a	a^2	f	fa^2
17	+3 8	14 44	2	28 88
15	+1 8	3 24	1	3 24
13	-0 2	04	3	12
11	-2 2	4 84	4	19 36
				51 60
				Σfa^2

Grouping Deviations as a Short Cut—Some saving in time and effort can be afforded in the solution of the standard deviation in data like those in Table 13, if we group them as in Table 14. Since the same measurement is repeated several times and its deviation from the mean

is the same every time, and also its deviation squared, we need to find the deviation and its square only once and multiply each x^2 by its frequency. The last column of Table 14 contains the fx^2 products, and it will be seen that their sum is again 51.60, from which the standard deviation will be the same as before.

A similar treatment may be given all grouped data, in which we let the midpoint of each interval be the X for all cases within the interval, and this X minus M gives us the deviation of all cases within the interval. From here on, the procedure is the same as that in Table 14. We shall not illustrate the steps by means of a special problem, for there are more efficient ways of dealing with grouped data.

The Standard Deviation by the Short Method.—The short method, which was employed in the preceding chapter to calculate a mean (pages 31ff), will now be extended in order to compute a standard deviation. The first steps are carried out exactly as previously to the point of finding the mean. The mean itself need not be known (since we are dealing with a guessed mean), but the correction is required, as will be seen in the following formula:

$$\sigma = i \sqrt{\frac{\Sigma fx'^2}{N} - c'^2} = i \sqrt{\frac{\Sigma fx'^2}{N} - \left(\frac{\Sigma fx'}{N}\right)^2} \quad (9)$$

where i = size of class interval.

x' = deviation from the guessed mean in terms of the class interval as the temporary unit.

c' = correction in the guessed mean, also in terms of the class interval as the unit.

The procedure is illustrated in Table 15, which is similar to Table 8 through column (4). For all class intervals, we need to know the fx'^2 products, and these are given in column (5). In each row, the fx'^2 product is found by multiplying the corresponding numbers in columns (3) and (4); *i.e.*, the first one, 25, is the product of 5×5 ; the second one is the product of 4×4 ; and the third, the product of 3×9 ; etc. We could, of course, square all the x' values first, then multiply by f in each interval, but we already have the fx' products that were needed to find the mean, and the x' values are already listed, and $fx'^2 = x'$ times fx' .

Next we sum the fx'^2 products to obtain $\Sigma fx'^2$. In Table 15, this is 230. To find c' , we divide $\Sigma fx'$ by N . In this case, it is $-24/50$, which equals -0.48 . We need c'^2 , which is 0.2304. Now, to apply formula (9), we need next to divide $\Sigma fx'^2$ by N , or $230/50$, which equals

TABLE 15 — CALCULATION OF THE STANDARD DEVIATION USING THE SHORT METHOD
(GUESSED-MEAN PROCEDURE)

(1) Score	(2) <i>f</i>	(3) <i>x'</i>	(4) <i>fx'</i>	(5) <i>fx'²</i>
55-59	1	+5	+ 5	25
50-54	1	+4	+ 4	16
45-49	3	+3	+ 9	27
40-44	4	+2	+ 8	16
35-39	6	+1	+ 6	6
30-34	7	0	0	0
25-29	12	-1	-12	12
20-24	6	-2	-12	24
15-19	8	-3	-24	72
10-14	2	-4	-8	32
	50 <i>N</i>		-24 $\Sigma fx'$	230 $\Sigma fx'^2$

$$c' = \frac{\Sigma fx'}{N} = \frac{-24}{50} = -.48$$

$$\sigma = 5 \sqrt{.280\%_0 - (-.48)^2} = 5 \sqrt{4.6 - .2304} = 5 \sqrt{4.3696} = 5 \times 2.09 = 10.45$$

4.6. Deduct from this c'^2 , or $4.6 - 0.2304$, and we have 4.3696. The square root of this is called for next, and this is 2.09. The last step is to multiply by i , the size of class interval; 2.09×5 equals 10.45, which is the standard deviation we have been seeking. We may now say that about two-thirds of the individuals should be expected between the mean minus 10.45 and the mean plus 10.45. Since the mean is 29.6, these limits are 19.2 and 40.0. Fortunately, for the sake of checking on this conclusion, these limits are close to the division points between class intervals (see Table 15). The four intervals included within these limits have in them 31 cases altogether, which are 62 per cent of the whole group. This is a little short of two-thirds but not unreasonably so.¹

¹ The probable error of a distribution is derived directly from the standard deviation by the formula $PE = .6745\sigma$. It is numerically about two-thirds as large as σ , as suggested by the ratio $.6745$. One more multiplication is required, and it is not quite so easily computed as the standard deviation. Its chief virtue is that in a normal distribution, 50 per cent of the measurements lie between the points at $-1PE$ and $+1PE$ from the mean. The writer has come to feel that this is not sufficient excuse for the inclusion of one more statistic to the already lengthy list,

Checking the Solution of the Standard Deviation—This kind of comparison is a rough check for the correct solution of the standard deviation. If the actual percentage of cases between plus 1σ and minus 1σ deviates too far from 68 per cent, there is probably something wrong with the calculation, and a recalculation is in order.

Another rough check is to compare the standard deviation obtained with the total range of measurements. In large samples ($N = 500$ or more) the standard deviation is about one-sixth of the total range. Or, stated in other terms, the total range is about 6 standard deviations. In smaller samples, the ratio of range to standard deviation becomes smaller, as indicated in Table 16.

TABLE 16—RATIOS OF THE TOTAL RANGE TO THE STANDARD DEVIATION IN A DISTRIBUTION FOR DIFFERENT VALUES OF N *

N	Range/ σ	N	Range/ σ	N	Range/ σ
5	2.3	40	4.3	400	5.9
10	3.1	50	4.5	500	6.1
15	3.5	100	5.0	700	6.3
20	3.7	200	5.5	1,000	6.5

* Adapted from Snedecor, G. W. *Statistical methods* P. 85. Ames, Iowa: Collegiate, 1937.

Since in the ink-blot data, $N = 50$, we should expect the range to be 4.5 times the standard deviation. The standard deviation 10.45 times 4.5 gives us an expected range of about 47 points. Actually the range was 46 points, which checks so closely as to give us confidence that our standard deviation is at least not grossly in error. It may seem strange that we use a less reliable statistic like range as a criterion of accuracy of a more reliable statistic like the standard deviation. The reasons are that (1) there can hardly be any error in computing such a simple thing as the range, whereas (2) there are chances of gross errors in calculating σ because of the many steps involved, for example, failing to make the final step of multiplying by i .

A Summary of Steps for Computing the Standard Deviation—The steps necessary for the calculation of σ by the short method are as follows:

particularly when the middle 50 per cent of the measurements can be more certainly delimited by the interval $Q_3 - Q_1$, the use of which does not assume normality of distribution

- Step 1. Complete Steps 1 through 6 already listed for finding the mean by the guessed-average route (see page 33).
- Step 2. Find for every class interval the fx' product. The most efficient way is to compute the product of x' times fx' for each interval. These products will all be positive.
- Step 3. Sum the fx'^2 products.
- Step 4. Divide this sum by N , carrying to four decimal places.
- Step 5. Find c'^2 , to four decimal places
- Step 6. Deduct the number found in Step 5 from that found in Step 4
- Step 7. Find the square root of the number found in Step 6, keeping two decimal places ¹
- Step 8. Multiply this number by the size of class interval. If N is large, save two decimal places; if small, round to one decimal place
- Step 9. Interpret the standard deviation in terms of the two-thirds principle.
- Step 10. Apply the rough check of comparing σ with the range and using the ratios of Table 16.

The Standard Deviation from Original Measurements.—If the number of measurements is not large, if the measurements themselves are small numbers, particularly when a good calculating machine is available, the best procedure for computing a standard deviation is by means of the formula

$$\sigma = \frac{1}{N} \sqrt{N \Sigma X^2 - (\Sigma X)^2} \quad (10)$$

in which the essential steps are:

- Step 1. Square each score or measurement.
- Step 2. Sum the squared measurements to give ΣX^2 .
- Step 3. Multiply ΣX^2 by N to give $N \Sigma X^2$.
- Step 4. Sum the X 's to find ΣX .
- Step 5. Square the ΣX to find $(\Sigma X)^2$.
- Step 6. Find the difference $N \Sigma X^2 - (\Sigma X)^2$.
- Step 7. Find the square root of the number found in Step 6.

¹ In this, in following steps, it is assumed that we are dealing with integral measurements. If they are in terms of decimal fractions or multiples of 10 or 100, this rule applies only after making the necessary allowance for the place of the decimal point

Step 8. Divide the number found in Step 7 by N (or multiply it by $1/N$).

On the calculating machine, the X 's and the X^2 's can be accumulated at the same time according to instructions provided with the machine. In tabular form, the solution of this kind is illustrated in Table 17.

TABLE 17—CALCULATION OF THE STANDARD DEVIATION FROM THE ORIGINAL MEASUREMENTS AND UNGROUPED DATA

X	X^2
13	169
17	289
15	225
11	121
13	169
17	289
11	121
13	169
11	121
11	121
132	1,794
ΣX	ΣX^2

$$\begin{aligned}
 \sigma &= \frac{1}{10} \sqrt{10(1,794) - 132^2} \\
 &= \frac{1}{10} \sqrt{17,940 - 17,424} \\
 &= \frac{1}{10} \sqrt{516} \\
 &= \frac{22.7}{10} \\
 &= 2.27, \text{ or } 2.3
 \end{aligned}$$

Grouping Original Measurements.—If the scores are conveniently grouped and their frequencies tabulated, as in Table 18, some saving in work can be effected. The steps by which we arrive at ΣfX and

TABLE 18—CALCULATION OF THE STANDARD DEVIATION FROM THE ORIGINAL MEASUREMENTS, WITH GROUPING

X	f	fX	X^2	fX^2
17	2	34	289	578
15	1	15	225	225
13	3	39	169	507
11	4	44	121	484
	10	132		1,794
	N	ΣfX		ΣfX^2

ΣfX^2 should now be easy to follow by an analogy to the last previous solution. Once those values are obtained, Steps 6 to 8 above can be followed to arrive at σ .

THE AVERAGE DEVIATION

The average deviation, or AD, is the arithmetic mean of all the deviations when we disregard the algebraic signs. We can disregard signs in this instance for the reason that we are not concerned about the *direction* of the deviations but only about their size. We treat them as if they were all positive. In terms of a formula

$$AD = \frac{\Sigma |x|}{N} \quad (11)$$

where AD = average deviation.

$|x|$, with the vertical bars embracing it, = absolute value of x , *i.e.*, disregarding algebraic sign.

To illustrate the solution of an average deviation, consider Table 19. The sum of the absolute deviations is 18.8. Divided by N , this gives 1.88 as the average deviation. Because of the small size of N , we should round to one decimal place and give the AD as 1.9.

TABLE 19—CALCULATION OF THE AVERAGE DEVIATION IN UNGROUPED DATA
(Mean = 13.2)

X	$ x $
13	0.2
17	3.8
15	1.8
11	2.2
13	0.2
11	2.2
17	3.8
13	0.2
11	2.2
11	2.2
	<hr/>
	18.8
	$\Sigma x $

$$AD = \frac{18.8}{10} = 1.88, \text{ or } 1.9$$

Relation of AD to σ .—The average deviation, in a normal distribution, is approximately .8 as large as the standard deviation, and so

the latter is about 1.25 times the average deviation. Within a range of plus and minus one average deviation around the mean are to be expected about 58 per cent of the cases. In the present problem, that range (the mean is 13.2) is from 11.3 to 15.1. Allowing that one of the 11's comes above 11.3, with the three 13's and the one 15, half the cases, or 50 per cent, fall within the limits just specified, which, in view of the small N and the peculiarly shaped distribution, is not bad. Comparing the AD with the standard deviation for the same distribution, we find that it is .84 as large, $1.88/2.27$, which is not far from the expected ratio of .8.

Short Cuts in Computation of the AD .—There are short procedures for computing an average deviation when data are grouped and when class intervals are employed. The use of the average deviation is so limited, however, that it is hardly worth the space it would take to describe those shorter methods here or worth the student's time to master them. The average deviation finds its greatest usefulness in distributions where N is rather small, and in these instances the procedure suggested is quite convenient. The only improvement to be mentioned in this connection is by way of grouping without using a class interval of more than one unit, such as is done in finding the standard deviation in Table 14. When data are grouped in class intervals larger than one unit, it is best to allow the midpoint of the interval to stand for that interval, find the actual deviation of this X value from the mean, and find the fx products, disregarding sign.

One thing about the average deviation that is different from the standard deviation is the fact that the deviations may be taken from the median rather than from the mean. In this case, if the distribution is skewed at all, the average of the deviations from the median will be smaller than the average of the deviations from the mean.

WHEN TO USE DIFFERENT MEASURES OF VARIABILITY

Several considerations come into the picture when we decide what measure of variability to employ in any situation. One is the reliability of the statistic; its relative constancy in repeated samples. In this respect, the statistics come in the order, from most reliable to least reliable: standard deviation, average deviation, semi-interquartile range, middle 80 per cent range, and total range. So far as quickness and ease of computation are concerned, the five are almost in reverse order to that just given. If further statistical computation is to be given the data, such as estimating reliability of the mean and of differences between means, computing coefficients of correlation, regression

equations, and the like, then the standard deviation is by all odds the one to employ.

As between standard deviation and average deviation, there is sometimes a choice. The standard deviation, because it derives from squared deviations, gives relatively more weight to extreme deviations from the mean. If a distribution should have an unusual number of extreme cases in one or both directions from the mean, the average deviation is a better statistic than the standard deviation. This rule includes cases of markedly skewed distributions.

The semi-interquartile range gives even less importance to extreme deviations than does the average deviation and would sometimes be given preference to both standard and average deviations for this reason. It gives more importance to the central mass of cases. When the median is the measure of central tendency adopted, Q should naturally be the companion measure of variability. Both are based upon the same principles. When distributions are truncated, only Q can justifiably be used to indicate variability.

THE COEFFICIENT OF VARIATION

Absolute versus Relative Variability.—It was said before that measures of variability are not directly comparable unless they are based upon the same scale of measurement with the same unit. It is even questionable whether one should compare absolute variabilities on the same scale when two groups have decidedly different means. For example, the variability in height of infants might naturally be expected to be less than the variability in height of adults. If we are interested in comparing the variability in height of infants, *as infants*, with variability in height of adults, *as adults*, we need to consider infant and adult norms. These norms are naturally given in terms of means or medians. We are here concerned with *relative* variability rather than *absolute* variability. The question is more correctly stated by saying, "Is the variability of infants' heights in ratio to their mean as great as the variability of adults' heights in ratio to their mean?" We therefore need to know the ratio of the standard deviation to the corresponding mean. It is customary to multiply this ratio by 100, which tells us what percentage of the mean the standard deviation is. The formula is

$$V = \frac{100\sigma}{M} \quad (12)$$

where V = *coefficient of variation*.

Relative Variability and Weber's Law.—One important application of the coefficient of variation is in the field of psychophysics. If we ask an observer to duplicate a 90-mm. line by free-hand drawing 50 times and if we then compute the mean and standard deviation of his reproductions, we may expect a mean something like 107 mm. and a standard deviation of about 5 mm. His coefficient of variation is 4.7; or, in other words, his variability is 4.7 per cent of his mean. In duplicating a line of 180 mm. 50 times, let us say that his mean is 195 mm. and his standard deviation is 8 mm. The variability has increased as well as his average. According to Weber's law, it should have kept in step with his increase in average and the coefficient of variation should be the same. V is now 4.1 per cent, or almost the same as before, but is perhaps lower than Weber's law requires. Results in the past have typically shown that with increasing mean, the absolute variability does not increase as rapidly in proportion, so that the relative variability decreases and does not remain constant, as according to Weber's law. We are not concerned here particularly with the validity of Weber's law except as it illustrates the importance of relative variability.

When Not to Apply the Coefficient of Variation.—One important word of caution is necessary concerning the application of V . It should not be applied unless we are rather certain that our measuring scale is one of equal units and, above all, unless the absolute zero point is taken into account. These qualifications almost entirely confine us to measuring scales with physical units, such as linear distances, weights, and time. They rule out ordinary test and examination scores, even mental-age and IQ units, and so materially reduce the areas of application of V in psychological investigations.

Exercises

1. Compute the middle 80 per cent ranges for distributions in Data D , E , and G .
2. Compute the interquartile and semi-interquartile ranges for the distributions in Data D , E , and G . Interpret your findings.
3. Compute the standard deviation for any or all of the distributions in Data C to G inclusive. Use any of the formulas that seem most convenient. Interpret your findings.
4. Compute the standard deviation in any or all of the distributions in Data H . Use any of the formulas that seem most convenient.
5. Compute the average deviation for any or all of the distributions in Data H .
6. Decide which measure of variability is wisest to employ with each of the distributions in Data C to G inclusive and which is second best. Give reasons.
7. In which of the same distributions would one be justified in computing a coefficient of variation and in which ones not? Give reasons.

8 Compute the coefficient of variation for each distribution in Data I. Interpret the table as it stands, and with your computed coefficients, also

DATA I—SCORES IN THREE MOTOR TESTS

Test	Tapping rate		Hand grip		Steadiness	
	Men	Women	Men	Women	Men	Women
Mean	210.4	184.0	42.1	23.9	5.64	5.13
Standard deviation	20.0	19.3	6.4	4.8	1.6	1.9
N ..	101	161	108	172	105	165

CHAPTER V

CUMULATIVE DISTRIBUTIONS AND NORMS

Many statistical procedures, particularly as applied to test scores, are based upon the cumulative frequency distribution. Heretofore we have given frequencies as belonging to certain scores or to class intervals. In this chapter, we are interested in the number of scores or measurements falling *below* a certain point on the measuring scale. The cumulative frequency corresponding to any class interval will be the number of cases within that interval *plus all those in intervals lower on the scale*.

CUMULATIVE FREQUENCIES AND CUMULATIVE DISTRIBUTION CURVES

How to Find the Cumulative Frequencies.—The cumulative frequencies are very readily found from the ordinary noncumulative frequencies. Our first example is with the already familiar ink-blot test scores (see Table 20). We list the scores in the first column just

TABLE 20—CUMULATIVE FREQUENCY DISTRIBUTION FOR THE INK-BLOT TEST DATA

(1)	(2)	(3)	(4)
Scores in the intervals	Exact upper limit of the interval	<i>f</i> Frequencies	<i>cf</i> Cumulative frequencies
55-59	59.5	1	50
50-54	54.5	1	49
45-49	49.5	3	48
40-44	44.5	4	45
35-39	39.5	6	41
30-34	34.5	7	35
25-29	29.5	12	28
20-24	24.5	6	16
15-19	19.5	8	10
10-14	14.5	2	2

as before, with high scores at the top, giving in column (1) the score limits of the class intervals. We next want a single score value to assign to each interval. Where before we used the midpoint, now we

choose the exact upper limit. The reason is that the frequency to be given corresponding to it will be all the cases *within* the class and *below* it. All those cases fall below the exact upper limit of the class. In column (3) are given the ordinary frequencies and in column (4), the cumulative frequencies. The cumulation is started at the bottom of the list in column (3). Below the upper limit of the lowest interval (14.5) are 2 cases. Below the upper limit of the second interval (19.5) are these 2 plus the 8 in the second interval, giving 10 as the cumulative frequency. In the third interval, we find 6 cases to add onto what we

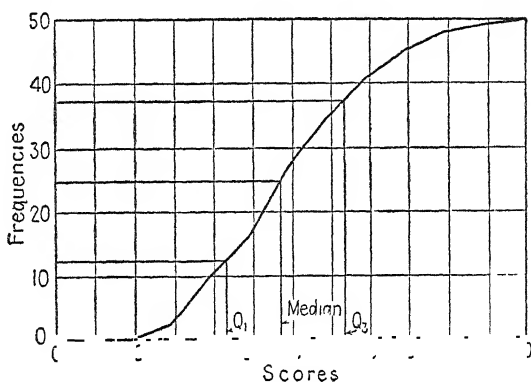


FIG. 8—A cumulative frequency distribution curve for the int-blot test scores

already have, making 16 for the third interval. And so it goes, each cumulative frequency being the sum of the preceding one and the frequency in the class interval itself. This continues until the last (top) interval is reached. The last cumulative frequency should be equal to N (here it is 50), if not, some error has been made.

Plotting the Cumulative Distribution.—Figure 8 shows the cumulative frequencies we have just obtained in Table 20 plotted against the corresponding scores (exact upper limits). The plotting here follows much the same routine as prescribed in Ch. II except that here we never plot the histogram form, only the type that connects neighboring dots with straight lines. Obviously we do not obtain a polygon but rather an S-shaped curve. In order to bring the curve to the base line at the left, we assume that a zero frequency comes at the lower limit of the bottom class interval (which is the same as the top of the interval just below it). As before, the total figure is about 60 to 75 per cent as high as it is wide.

Determining Quartiles Graphically.—It is of interest to point out here the ease with which the quartiles can be graphically determined

or read off the curve in Fig. 8. To find the median (Q_2), we first locate the frequency of 25 ($N/2$) on the vertical axis. Draw a horizontal line over to the curve at this level. At the point where it intersects the curve, drop a perpendicular to the base line. Where this cuts the base line, read the score value. On ordinary graph paper, Q_2 can be read accurately to one decimal place. Q_1 would be similarly determined at the level of 12.5 on the frequency scale and Q_3 , at the level of 37.5.

Distribution of Cumulative Percentages and Proportions.—Previously we have had reason to transform frequencies into percentages for the sake of comparing two distributions where N differs. The same reason, plus more important ones, prompts us more frequently to transform cumulative frequencies into percentages. In Table 21, another example of cumulative frequencies is given. They are obtained here [column (4)] just as before. We now wish to find what percentage of 86 each cumulative frequency is. The arithmetic is simply a matter

TABLE 21—CUMULATIVE FREQUENCIES, PERCENTAGES, AND PROPORTIONS FOR MEMORY-TEST SCORES

(1) Scores	(2) X	(3) f	(4) cf	(5) Cumulative %	(6) cp
41-43	43.5	1	86	100.0	1.000
38-40	40.5	4	85	98.8	.988
35-37	37.5	5	81	94.2	.942
32-34	34.5	8	76	88.4	.884
29-31	31.5	14	68	79.1	.791
26-28	28.5	17	54	63.0	.630
23-25	25.5	9	37	43.0	.430
20-22	22.5	13	28	32.6	.326
17-19	19.5	8	15	17.4	.174
14-16	16.5	3	7	8.1	.081
11-13	13.5	4	4	4.7	.047
8-10	10.5	0	0	0.0	.000

of multiplying each cumulative frequency by $100/N$. This fraction, $100/86$, is equal to 1.1628. It is well here to keep a liberal number of decimal places. In Table 21, the cumulative percentages in column (5) are obtained by multiplying each frequency in column (4) by 1.1628. These need not be given to more than one decimal place. Sometimes it is preferable to work in terms of cumulative *proportions*, which are given in column (6). Whereas with percentages the base is

100, with proportions the base is 1.00. Each proportion is therefore simply 1/100 of the corresponding percentage. The reason for using proportions will be explained later, here we shall be concerned with percentages.

The Cumulative Percentage Curve, or Ogive.—In Fig. 9, the cumulative percentages we have just obtained in Table 21 are plotted as points against the corresponding score points (exact upper limits of class intervals). Again, an S-shaped curve results. Now that it is standardized as to height, it is sometimes called an *ogive*¹. The *ogive* is,

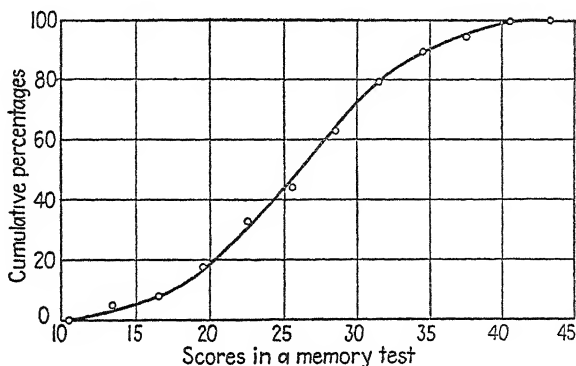


FIG. 9—Smoothed cumulative distribution curve for the memory-test scores. Frequencies are in terms of percentages. This type of curve is called an *ogive*.

in other words, *the cumulative percentage distribution curve*. Two ogives are much more readily compared than two ordinary cumulative curves because of their common height. But this is not the only use of an ogive, as we shall soon see.

CENTILE NORMS

Finding Centiles by Interpolation.—A *centile* (often superfluously called *percentile*) is a point on the scoring scale below which are any given percentage of the cases. For example, the 90th centile is the point below which are 90 per cent of the scores, and the 24th centile is the point below which are 24 per cent of the scores.

Deciles and Tenths.—We have already seen how to interpolate in order to compute a median and the other quartiles. Actually, the median is at the 50th centile, Q_1 is at the 25th centile, and Q_3 is at the 75th centile. It is but a step further to generalize this to any centile

¹ The reader will frequently find that the ogive is presented with the percentages on the horizontal axis and the score scale on the vertical axis. The writer prefers to keep the diagram here consistent with that for noncumulative distributions.

one desires. We could choose to interpolate any centile, the 63d, the 81st, or the 8th. Our interest in testing happens to center upon the centiles that are multiples of 10—the 90th, 80th, 70th, etc., down to the 10th. These are called the *deciles*, for they divide the distribution into tenths, just as the quartiles divide it into quarters and the median, into halves.

The Process of Interpolation—The principle of interpolating is not new. But in Table 22 is shown how we may work out the deciles

TABLE 22—CALCULATION OF CENTILES BY INTERPOLATION IN THE MEMORY-TEST DATA

Percentage below the centile point	Number of cases below the centile point	Cumulative frequency actually below the interval containing the centile point	Lower limit of interval containing the centile point	Distance of centile point above lower limit	The centile point
90	77.4	76	$34.5 + \frac{14 \times 3}{5}$		35.3
80	68.8	68	$31.5 + \frac{8 \times 3}{8}$		31.8
70	60.2	54	$28.5 + \frac{62 \times 3}{14}$		29.8
60	51.6	37	$25.5 + \frac{146 \times 3}{17}$		28.1
50	43.0	37	$25.5 + \frac{6 \times 3}{17}$		26.6
40	34.4	28	$22.5 + \frac{64 \times 3}{9}$		24.6
30	25.8	15	$19.5 + \frac{108 \times 3}{13}$		22.0
20	17.2	15	$19.5 + \frac{22 \times 3}{13}$		20.0
10	8.6	7	$16.5 + \frac{16 \times 3}{8}$		17.1

systematically. The complete headings of the table make the work almost self-explanatory, but let us follow through one or two examples. First we need to know how many cases out of the total of 86 we need to include in any given percentage. Ninety per cent of 86 is 77.4, which

we find in column (2). We must count up the scoring scale among the frequencies until we include 77.4 cases. Reference to Table 21 shows that we get by accumulation 76 cases up to the score point 34.5. We need 1.4 more cases among the 5 in the next higher interval. There are 3 score units in the interval; so we have to proceed $1.4/5$ times 3, or, as given in column (5) of Table 22, we add to 34.5 the amount $\frac{1.4 \times 3}{5}$, which gives us 35.3 as the centile point. We say that P_{90} (90th centile) equals 35.3. To take a second example, let us solve for P_{10} . Ten per cent of 86 is 8.6. Counting up to a score point of 16.5, we find 7 cases, which leaves us needing 1.6 more out of the 8 in the next interval. P_{10} is therefore equal to $16.5 + \frac{1.6 \times 3}{8}$, which = 17.1. The remaining centiles are similarly determined and are listed in the last column of Table 22.

The Utility of Centile Norms.—Test scores of various kinds are frequently interpreted in terms of centile norms, for very good reasons. In the first place, a raw score of so many points means very little to us. Tell a student's adviser that his advisee made a score of 59 points in an algebra-achievement examination, 175 points in an English-achievement examination, and 121 points in a general scholastic-aptitude test, and without further information the adviser does not know whether his advisee is low in all tests, high in all tests, or low in one or two and high in the remaining. But tell him that a score of 59 points in algebra is at the 99th centile, the 175 points in English is at the 32d centile, and the 121 in scholastic aptitude is at the 48th centile, when those centiles were established by the scores from 1,500 freshmen entering the University with the advisee in question; then he will have some usable information. The student in question is extremely high in algebra, moderately low in English, and about average in general scholastic aptitude. The chief utility of centile norms is (1) to give some conception of the general level of a score in a known population, and (2) to put scores from different tests on a comparable basis.

Finding Centile Norms by Interpolation.—If we wished to have a table of centile norms for the memory test, we could now use the nine decile points already found by interpolation as they are listed in the last column of Table 22. Then when a student came along with a score of 22 we could say that he is at the 30th centile; another student with a score of 30 is at the 70th centile, etc. When a score came up that is not exactly listed we could find its centile value by interpolation. For

example, a score of 21 would be at the 25th centile, and a score of 27 would be at about the 53d centile.

Centile Norms from Smoothed Ogives.—But there are objections to be made to the use of interpolated centiles as norms. Chance irregularities in distribution from a small sample often give a distorted picture of the true situation that probably obtains in the larger population. After all, it is the larger population that we wish to represent in our norms, or at least we should like to compare future individuals' scores with something more stable and general than our limited sample. For this reason the writer recommends that centile norms be set up in terms of the smoothed ogive. Interpolated norms are derived from the unsmoothed curve and, as was said, they are affected by minor irregularities that are probably a peculiarity of this sample only and not of the general population. The smoothed ogive may be taken as an estimation of the distribution of the general population of which our group is a sample. When a sample is large, very little smoothing is necessary. Even with small samples at times surprisingly little smoothing need be done.

In Fig. 9, a smoothed ogive (by inspection and free-hand drawing) has been drawn. The aim is to bring it as close as possible to all points, and if points must be untouched by the curve, there should be about as many below the curve as above it. If too glaring discrepancies occur between points and curve after smoothing, it is probably best to discard the attempt to use these data as a basis for norms or else to add more cases until sampling fluctuations are greatly reduced.

Reading Centile Scores from a Graph.—Having satisfied oneself as to the smoothed ogive, the next step is to read off the diagram the score points corresponding to the centile points for which norms are required. For this purpose the diagram should be enlarged sufficiently for easy reading and the graph paper finely ruled so that score points may be accurately read to one decimal place. In Table 23 are given the score points corresponding to centiles 10 to 90, as before, but also to 95 and 99 at the upper end and to 5 and 1 at the lower end. The reason for including these extra points at the extremes is that there is actually a great range of ability above the 90th centile and also below the 10th centile.

A Defect in Decile Scales.—One defect of the centile scale, as a measuring scale, is that it exaggerates individual differences relatively near the center of the distribution as compared with the ends. Giving score norms corresponding to the centiles beyond 10 and 90 com-

TABLE 23 —CENTILE NORMS FOR THE MEMORY TEST, DERIVED FROM THE SMOOTHED OGIVE

Centile	Score point	Integral score
99	40 5	41
95	37 1	38
90	34 9	35
80	31 8	32
70	29 5	30
60	27 9	28
50	26 1	27
40	24 3	25
30	22 5	23
20	20 4	21
10	17 5	18
5	14 9	15
1	11 9	12

pensates for this defect to a large extent. Because of this same defect, it is not the best practice to work with decile norms, for to do so often leads the user of the norms to lay too much stress upon differences among the great average group and too little upon those where tests discriminate best. It is probably best that decile norms be consigned to the limbo of forgotten procedures. In their place the writer strongly urges the employment of a kind of profile chart that some testers have already put into use. This kind of chart will be described shortly.

Integral Centile Points.—Before doing that, however, a further word of explanation of Table 23 is in order. The last column of "integral scores" is merely a revision of the second column by way of rounding to whole numbers. Tables of norms are frequently given in terms of whole numbers, mainly because scores are obtained as whole numbers. We should say that an obtained score of 41 is better than 99 per cent of the group can make, and a score of 18 is better than only the lowest 10 per cent can make. It should be noticed that every fractional score is rounded upward to the next whole number; thus 37.1 becomes 38. Since an obtained score of 37 covers a range of 36.5 to 37.5, more than half of those making this score would *not* be better than 95 per cent. The first score, counting from below upward, that is totally better than 95 per cent is a score of 38. This is why in this and in other cases we round upward to the next higher integer.

A Graphic Profile Chart.—Many profile charts based upon centiles show graphically the deciles at equidistant levels along the scale. This gives an erroneous conception of the relative spacing of ability or talent, as was pointed out in a preceding paragraph. Actual differences in ability are probably more accurately indicated by the raw-score units than they are by centile units, which relatively magnify the central portions of the distribution. If it is assumed that the actual distribution for the norm group is Gaussian or normal in shape, the relative spacing of the various centiles that we customarily include in our norms should be as given in Table 24. In the first column are the cus-

TABLE 24.—THE DISTANCE OF CENTILES FROM THE MEAN IN NUMBER OF STANDARD DEVIATIONS IN A NORMAL DISTRIBUTION

Centile	Number of Sigma from the Mean
99	+2 33
95	+1 64
90	+1 28
80	+0 84
70	+0 52
60	+0 25
50	0 00
40	-0 25
30	-0 52
20	-0 84
10	-1 28
5	-1 64
1	-2 33

tomary centiles. In the second column are the corresponding distances from the mean (and median) when the standard deviation of the distribution is adopted for convenience as the unit. The correspondence of deviation from the mean with centile depends entirely upon the mathematical relations that hold true for the normal distribution curve and the reasons for this need not concern us here. The writer merely proposes to use this spacing of the centiles in setting up a profile chart and has done so in Fig. 10.

Here, in Fig. 10, each centile is drawn at a distance from the mean proportional to its corresponding sigma distance given in Table 24; *i.e.*, centiles 99 and 1 are 2.33 units from the mean, centiles 90 and 10 are 1.28 units away, etc., though those units are not labeled in the chart and need not be. Once having located them at the proper distances, we may forget the sigma distances.

Provision has been made for four tests in the profile chart, the memory test whose norms we have determined in previous parts of this chapter, a vocabulary test, a word-building test, and a sentence-construction test, whose norms were determined elsewhere. For the memory test, the integral scores have been written in at their cor-

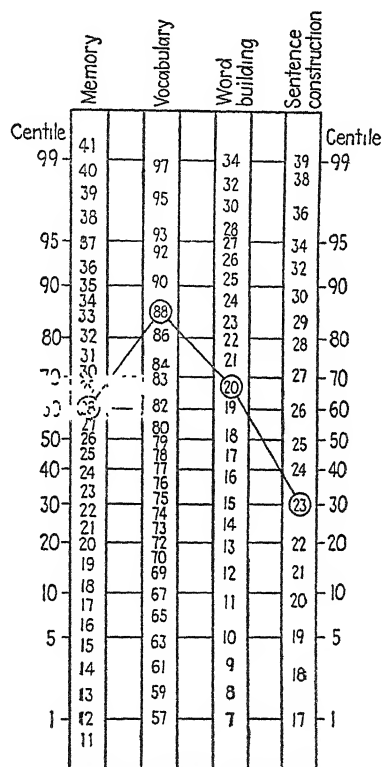


FIG. 10—An example of a profile chart based upon centile norms. Note that the centiles are not spaced at equidistant intervals but at intervals based upon corresponding sigmas from the mean (see Table 24)

responding centiles, being guided by the list of score points in column (2) of Table 23. Once the scores nearest those points are located and written in the diagram, the other, intervening scores can be introduced. The same was true for the other test norms, though because of crowding, some integral scores have been omitted. The student whose profile is shown earned raw scores of 28, 88, 20, and 23, respectively, in the four tests. Those four scores have been encircled and then connected with straight lines to complete the profile. We can

now see at a glance the general trend of this student's ability in these four tests taken together, and we can read off his centile rating in each test at a glance. Furthermore, a much more accurate conception of his fluctuation in ability is given than would have been true in a diagram with equidistant deciles.

A Bar Diagram of Distributions of Scores.—A relatively new graphic device for picturing distributions of scores is shown in Fig. 11.¹ The bar diagrams there illustrate the distributions of three groups of students who were taught by three different instructors but who were

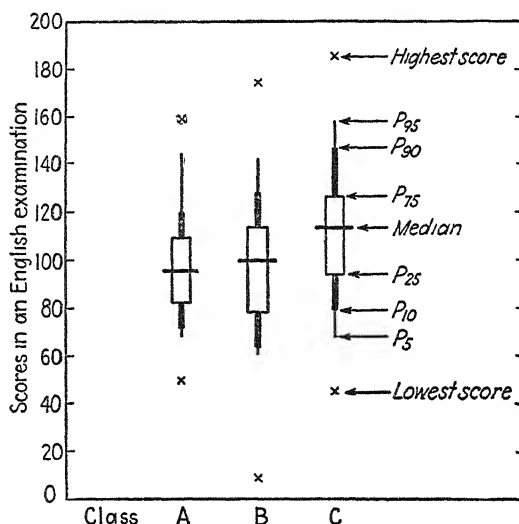


FIG. 11.—A graphic device for visual comparison of distributions, showing important points.

given the same final examination, an objectively scored achievement test in English. The median of each group is marked by a short horizontal line through the bar at the median-score level. The range of the middle 50 per cent (from P_{25} to P_{75} , or from Q_1 to Q_3) is shown in each case by the open rectangle. The black bars extend out to the points P_{10} and P_{90} —in other words, to include the middle 80 per cent of the cases. The lines extend to points at P_5 and P_{95} , or to include the middle 90 per cent of the cases. The highest and lowest single scores are marked by the small x 's. Thus several meaningful centile points are labeled as well as the entire range.

¹ Similar diagrams have been in use for some time by the Cooperative Test Service.

Interpretation of Bar Diagrams.—One important use of bar diagrams is the ready comparison of groups that they afford. In Fig. 11, for example, it is obvious that the three medians come in the order 1, 2, 3 for groups *C*, *B*, and *A*, respectively. The variabilities of the three groups come in the order *B*, *C*, and *A* when we depend upon total ranges. The groups come in about the same rank order for variability when we compare ranges of middle 90 per cent, but again the order *B*, *C*, *A* is probably correct in comparing middle 50 per cents, though *B* and *C* are very close together in this respect. As to topmost scores, they come in the same order as for medians *C*, *B*, *A*, but for bottom scores the order is *A*, *C*, *B*. As to skewness, the most symmetrical distribution, all things considered, is probably that for group *B*, and the least symmetrical is for group *A*, which is positively skewed. The special virtue of this kind of comparison, as contrasted with that afforded by means of frequency polygons and ogives, is that many more facts about a distribution can be recorded, and yet because of no overlapping of the drawings there is direct comparison without confusions.

Exercises

1. Carry through the following steps for the first distribution of chemistry-aptness scores in Data *C* (page 27):
 - a* Find the cumulative frequencies, and tabulate them
 - b* Plot a cumulative distribution curve similar to Fig 8 (page 65)
 - c* Find the cumulative percentages and proportions, and tabulate them
 - d* Plot the ogive distribution, showing the smoothed curve.
 - e* Compute the interpolated centiles that divide the distribution into tenths
 - f* Derive centile norms from the smoothed ogive, and set up a table of norms
 - g* Prepare a centile profile chart including the norms for this test and any others for which you have norms.
- 2 Repeat the steps, particularly Steps *a*, *c*, *d*, and *f*, for any other distribution of test scores
3. Prepare bar diagrams like those in Fig 11 (page 74) for comparing two or more distributions, such as the two in Data *C* or Data *G*.

CHAPTER VI

THE NORMAL DISTRIBUTION CURVE

Repeatedly have sets of measurements in psychology and education yielded frequency distributions that resemble the bell-shaped normal, or Gaussian, curve. And because the normal curve has so many useful mathematical properties, it is quite natural that we should exploit those properties in dealing with psychological and educational data. Without the use of the Gaussian curve and its convenient characteristics, many things that we now do with data would otherwise be impossible. It is important, therefore, that the student develop at least a moderate understanding of the normal curve in order that he may wisely apply the statistical procedures that depend upon it.

Normality of Distribution Is Assumed.—It must be confessed at the outset that no set of data ever obtained, whether they be measurements of a group of individuals with respect to some biological, psychological, social, or educational trait or whether they be repeated observations of a single phenomenon, ever conforms exactly to the normal distribution pattern. Even though the larger population from which our sample came is perfectly normally distributed (even this is probably never strictly true), sampling, no matter how extensive or representative it may be, is bound to give us some irregularities, with deviations from the normal form. Whenever, therefore, we treat our data as if they were normally distributed, or arose from a population that is normally distributed, we are assuming an ideal pattern for the sake of simplicity, rationality, and convenience. Sometimes we are more justified and sometimes less; we can never be absolutely sure, because the entire population is rarely or never measured, and the true shape of distribution is never known.

We can justify our assumption of normality in several ways. One is the rational approach, which attempts to point out that the phenomenon we are measuring results from a number of independent causes occurring in chance combination, as in the tossing of coins or in the combinations of hereditary genes. Very rarely is this kind of argument possible because of our ignorance of underlying causes. Another kind of approach is empirical, in which we can show that, with

the use of the measuring scale that we did use, the grouped data present a frequency distribution that obviously possesses a bell-shaped contour. Furthermore, there are statistical tests that can be applied to show whether or not the frequencies we obtained deviate too much from the normal-curve picture to cause us to reject our hypothesis that the data came by random sampling from a normally distributed population.

Two Reasons for Caution.—There are two considerations, however, which should cause us to pause before making the hypothesis or assumption of normality. One has to do with the question of sampling and the other with the question of the validity of our measuring scale. A population may well be normally distributed, yet because of our method of drawing cases for measurement we may obtain a skewed or otherwise distorted form of distribution. This is a case of *biased sampling*. A large population of ten-year-old children would probably be distributed normally when measured for mental age. But if we confine ourselves to ten-year-old children in the fourth grade only, where most ten-year-olds are probably present because of mental retardation and a few for other reasons, the distribution of mental ages would be positively skewed. The ten-year-olds in the sixth grade would probably yield a negatively skewed distribution, for the majority of them are accelerated by reason of precocity and a few for other causes. Both are cases of biased sampling. An unbiased, representative sampling would not confine itself to fifth-grade children, but would take ten-year-olds in correct ratios from all grades where they appear, would take them in correct proportions as to sex, economic status, and other factors considered significant.

Another factor making for skewness even when the population is normally distributed, also, at other times, making for normality of distribution in the sample when the population distribution is skewed, is a systematic change in the size of unit of measurement over the range of cases measured. If in a mental test the measurement is the number of correct responses and if among the easy items there is little difference in ability required to pass additional items but among the hard ones there are great increments of ability involved, we have a “rubbery” yardstick with small units at the lower end and large ones at the upper end. If on a scale of this ability where *ideal* units are equal, there is a genuinely normal distribution, what will happen to the distribution on our “rubbery” scale? The effect of the faulty scale here would be to enlarge (apparently) differences at the lower end of the scale. Negative skewing would be the result, but it is an artificial

skewing. Had the true distribution of the population been positively skewed, the fault in the scale might have been such as just to correct it for skewness and to yield what is apparently a symmetrical, normal distribution.

These cautions kept in mind should serve to inhibit many dogmatic assertions that might otherwise be made about the shape of distribution of measurements. The shape of distribution is always a function of the kind of measuring scale, and all conclusions that involve form of distribution should take this fact into account. The conviction that general populations are genuinely normally distributed with respect to most qualities is very strong, however; so it is usually the marked deviation from normality in a sample that arouses questions. We may then question either our method of sampling or our measuring scale. One or both of these factors may be responsible for the discrepancy. But when our sample distribution turns out reasonably normal in appearance, because of the conviction just mentioned we may feel some assurance that our sampling and our measuring scale are probably free from distortions, though of course we can never be certain of this. The conviction does lead us to apply the Gaussian curve in many useful ways, even in turning crude judgments into scaled measurements, as we shall see later. We frequently feel that the risk in making the normal assumption is well worth while because of the invaluable results and conclusions it affords. We can always state our conclusions with the reservation that they are true to the extent that our assumptions are valid. As a matter of fact, all other conclusions should be couched in similar terms, for none is without its foundation of assumptions of one kind or another, whether stated or not. All scientific conclusions rest on assumptions, in the final analysis, and he who would know the import of those conclusions best is the one who knows those assumptions best.

THE NATURE OF THE NORMAL CURVE

The Relation of the Normal Curve to Probability.—The Gaussian curve is also sometimes called the *normal probability* curve and is said to be the result of the “laws of chance.” In a sense, this is true. We cannot here go into an involved discussion of probability and of the way in which the Gaussian curve is logically related to probability. It is sufficient for our present purposes to point out the usual example of how a normal distribution can be approximated by means of coin tossing. If we thoroughly shake a set of 6 coins and toss them to land where and how they may, the result can turn out in seven ways; the

number of heads can vary all the way from 0 to 6. In a total of 64 tossings, according to the principles of probability, we should expect the following frequencies for various numbers of heads:

Heads	0	1	2	3	4	5	6
Frequencies	1	6	15	20	15	6	1

If we tossed the 6 coins twice as many times, we should expect these frequencies to be doubled. Actually obtained frequencies will deviate from these expected ones by small amounts. In one such experiment with 128 tosses, the obtained frequencies were as given here:

Heads	0	1	2	3	4	5	6
Obtained frequencies	2	14	25	38	36	12	1
Expected frequencies	2	12	30	40	30	12	2

This situation is shown graphically in Fig 12, where the obtained frequencies furnish the basis for the histogram and the expected frequencies furnish the basis for the superimposed normal curve.

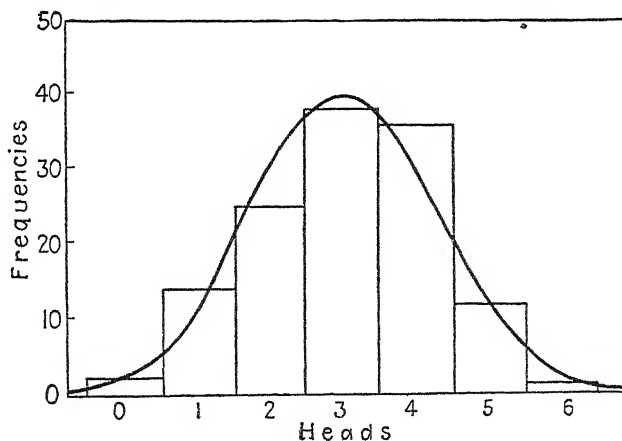


FIG. 12.—A distribution curve representing the frequencies with which various numbers of heads are expected by chance in tossing six coins, also, in histogram form, the obtained frequencies of 128 tossings

A 6-coin problem gives us a 7-sided frequency polygon (not counting the base line). A 10-coin problem gives us an 11-sided contour, etc., the number of sides being equal to the number of coins plus 1. If we do not enlarge the base line of our distribution but keep subdividing it into smaller and smaller units as we increase the number of coins,

the contour of the distribution curve approaches the smooth bell form. The number of class intervals we choose in grouping obtained measurements has nothing to do with the number of coins, our choice being entirely arbitrary. The class intervals and their frequencies merely give us descriptions of the contour at points along the way. If there are things like coins in the phenomenon we are measuring (*i.e.*, "coins" such as genes, which may be present or absent, or such as responses that do or do not occur) we almost always lack information as to how many such "coins" are operating. Probably there are a great many, although even if there were only 6, as in the coin example, and if our measurements naturally fell therefore into seven class intervals, the normal distribution could still be roughly approached, as can be seen in Fig 12.

The Equation for the Normal Curve.—Mathematically, when we are dealing with the properties of the normal curve, it is the situation with an infinite number of "coins" that we suppose. This enables the mathematician to give to the curve an equation that describes the relationship of a frequency to its corresponding measurement. This equation reads

$$Y = \frac{N}{\sigma \sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}} \quad (13)$$

where Y = frequency.

N = number of measurements.

σ = standard deviation of the distribution.

π = 3.1416.

e = 2.718 (the base of the Naperian system of logarithms).

x = deviation of a measurement from the mean (or $X - M$).

Since the values for π and e are known, if we substitute them in the equation, it becomes

$$Y = \frac{N}{2.5066\sigma} 2.718^{\frac{-x^2}{2\sigma^2}}$$

For any distribution we may have at hand, we know the values for N and for σ , and these can be inserted in their places in the equation. The equation would then be in a form with only Y and x the unknowns. We could then assign certain values to x , within the range of our measurements, and then solve the equation for the corresponding values of Y . In this way, we could determine the entire normal distribution curve that best fits our data. The arithmetical work would be a little laborious, but fortunately we have the use of statistical

tables to aid us in this. Table B (see page 318) is one well suited to this purpose

Determining the Best-fitting Normal Distribution for a Set of Data.

For the sake of an illustration that will help us to appreciate the meaning of the normal curve, let us find the expected frequencies in a particular instance, a distribution of 86 scores in a memory test. The best-fitting normal curve for any set of data has the same mean and standard deviation as those computed from the actual data. The distribution of obtained frequencies of memory-test scores is given in column (7) of Table 25. The mean of this distribution is 26.1, and the

TABLE 25—OBTAINING THE EXPECTED FREQUENCIES f_e IN THE CLASS INTERVALS FOR THE MEMORY TEST, ON THE ASSUMPTION THAT THE TRUE DISTRIBUTION IS NORMAL

(1)	(2)	(3)	(4)	(5)	(6)	(7)
Scores	X Midpoint	x Deviation	z Standard score	y From Table B	f_e Expected frequency	f_o Observed frequency
44-46	45	+18.9	2.93	.0055	0.2	0
41-43	42	+15.9	2.47	.0189	0.8	1
38-40	39	+12.9	2.00	.0540	2.2	4
35-37	36	+9.9	1.53	.1238	5.0	5
32-34	33	+6.9	1.07	.2251	9.0	8
29-31	30	+3.9	0.60	.3332	13.3	14
26-28	27	+0.9	0.14	.3951	15.8	17
23-25	24	-2.1	-0.33	.3778	15.1	9
20-22	21	-5.1	-0.79	.2920	11.7	13
17-19	18	-8.1	-1.26	.1804	7.2	8
14-16	15	-11.1	-1.72	.0909	3.6	3
11-13	12	-14.1	-2.19	.0363	1.5	4
8-10	9	-17.1	-2.65	.0119	0.5	0
Sums					85.9	86.0

Each column of numbers is derived from the one preceding by the following computations (see text for explanations)

Column (3) $x = X - 26.1$

Column (4) $z = x/6.45$

Column (5) y comes from Table B

Column (6) $f_e = 40 \times y$

standard deviation is 6.45. Our task is to find the frequencies to be expected in the same class intervals for a normal distribution with a mean of 26.1, a standard deviation of 6.45, and an N of 86.

Standard Measurements or Scores.—In order to use equation (13) to find these frequencies, we must know how far each class interval deviates from the mean in terms of standard deviations. Each interval is given the value of its midpoint as its point on the score scale X . These X values are listed in column (2) of Table 25. Note that we have included one class interval beyond the range of obtained scores at each end of the distribution. This is because the best-fitting normal curve usually has some small frequencies (perhaps fractional) in those extreme positions, even though the obtained frequencies there are zero. The equation for the normal curve calls for *deviations* rather than original scores—in other words, for $X - M$, or small x , for each class interval. These are listed in column (3). In this problem, each one is found by the solution of $X - 26.1$ for every interval. A simple check is to see that each one is three units (the size of the interval) distant from its immediate neighbors. The next step involves a new process; the determination of the *standard measurement or standard score*, for every interval. The standard score is given by the formula

$$z = \frac{x}{\sigma} = \frac{X - M}{\sigma} \quad (14)$$

In the equation for the normal curve, it will be seen that the exponent of e , which is $-x^2/2\sigma^2$ can be written $-(1/2)(x/\sigma)^2$, or in other words, it is $1/2$ times the standard score squared. We shall find the standard score invaluable again and again. The statistical tables are constructed on the basis of standard scores. It matters not, then, what our original means and standard deviations are numerically. Reducing all raw scores to standard scores places them all on the same basis or common denominator. For our illustrative problem, the standard scores are given in column (4) of Table 25. Each number in column (4) is obtained by dividing the corresponding number in column (3) by 6.45, the standard deviation.

Determining Frequencies for the Class Intervals.—Having obtained the standard score for each class interval, we are now ready to look up the corresponding ordinate in the general statistical table, Table B. These are listed in column (5) of the work table. The ordinates in this table are not exactly the frequencies we have been wanting to find. Those frequencies also depend upon N [see equation (13)]. Table B is constructed on the assumption that $N = 1$, and $\sigma = 1$. For our distribution of 86 cases and a different σ , we must make a certain adjustment. We must multiply each y value by a certain number to

find the expected frequency f_e . The general formula is

$$f_e = \left(\frac{iN}{\sigma} \right) y \quad (15)$$

In this problem,

$$\frac{iN}{\sigma} = \frac{3 \times 86}{6.45} = \frac{258}{6.45} = 40.0$$

When this multiplier is used with the numbers in column (5), the frequencies we desired are finally forthcoming, and they are given in column (6).

Comparing Obtained and Theoretical Frequencies.—As a rough check upon all the work, we sum these frequencies, and the result should be very close to N but will usually be slightly less than N , because in the normal curve there are still fractions of frequencies even beyond the limits we have included here. Had we not gone one class interval beyond the obtained data, we should have lost .2 of a frequency at the upper end and .5 at the lower, and the sum would have been 85.2 instead of 85.9. As it is, we have still lacking only .1 of a case; not enough to worry about, and we may accept our check as one indication of correct work. A comparison of expected with obtained frequencies is always a rough check but is very rough, because we expect small discrepancies within class intervals. Looking down the columns, we find only one or two serious discrepancies. One is the difference between 15.1 and 9, and the other is between 1.5 and 4. Both the obtained frequencies of 9 and 4 are out of line but are probably merely chance discrepancies, coming under the heading "errors of sampling," and are no more serious than may be expected in a coin-tossing experiment.¹

Plotting the Best-fitting Normal Curve.—We could now use the expected frequencies as the basis of plotting the best-fitting, smooth, normal distribution curve for the memory-test data. If plotting such a curve is our only objective, however, we have done some unnecessary work. A shorter procedure for locating enough points for drawing the smooth best-fitting curve will now be explained. It follows precisely the same principles laid down in the previous discussion.

¹ The customary way of determining whether the discrepancies between theoretical and obtained frequencies are so large as not to be attributable to sampling errors is to employ the chi-square test (see Ch. IX, particularly p 173). The chi-square test, as applied to the normal-curve hypothesis, tells us the probability that an obtained set of frequencies is not normally distributed

But instead of being tied down to class intervals and their midpoints for our x values, we instead arbitrarily choose standard scores at convenient values $\frac{1}{2}\sigma$ apart, as in the first column of Table 26. Since they are simple numbers, no interpolation will be necessary in using Table B. Since the positive standard scores duplicate the negative ones, half the work of looking up y values is obviated, unless one wishes

TABLE 26—OBTAINING THE BEST-FITTING NORMAL CURVE FOR THE DATA ON THE MEMORY TEST FOR THE PURPOSE OF PLOTTING THE CURVE

(1)	(2)	(3)	(4)	(5)
z Standard score	y From Table B	f_e Expected frequency	x Deviation	X Raw score
+3 0	.0044	0 2	+19 4	45 5
+2 5	0175	0 7	+16 1	42 2
+2 0	.0540	2 2	+12 9	39 0
+1 5	1295	5 2	+ 9 7	35 8
+1 0	2420	9 7	+ 6 4	32 5
+0 5	3521	14 1	+ 3 2	29 3
0 0	3989	15 8	0 0	26 1
-0 5	3521	14 1	- 3 2	22 9
-1 0	2420	9 7	- 6 4	19 7
-1 5	1295	5 2	- 9 7	16 4
-2 0	0540	2 2	-12 9	13 2
-2 5	0175	0 7	-16 1	10 0
-3 0	0044	0 2	-19 4	6 7

The numbers in the columns are obtained as follows

Column (1). Arbitrarily chosen.

Column (3): $40 \times y$

Column (4) $6.45 \times z$.

Column (5). $x + 26.1$

to repeat the process as a check. The expected frequencies are again found by y by iN/σ , in this case, by 40. To find the score points on the scale of measurement corresponding to our expected frequencies, we require the last two columns. The deviation x in column (4) is found by the equation

$$x = z\sigma \quad (16)$$

which is derived from formula (14). Then the corresponding raw score in the last column is simply equal to $M + x$, or, in this problem, $26.1 + x$.

Having these score points and their corresponding frequencies, we can construct the graph shown in Fig 13. The observed frequencies (f_o) are also plotted as circlets to show where they fall with respect to the best-fitting normal curve. The reasonableness of the fit is rather obvious. It would probably have been not so easy to duplicate this

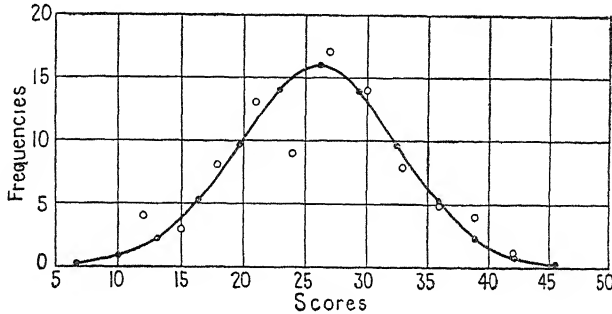


FIG 13 —The best-fitting normal distribution curve for the memory-test data. Obtained frequencies are represented by the circlets. The normal curve has the same mean and standard deviation as the obtained distribution.

normal curve by the smoothing process recommended in Ch II. We may say by way of general conclusion that if our obtained mean and standard deviation approximate closely the mean and sigma of the normally distributed population from which our sample came, the distribution for the population looks like the normal curve in Fig. 13.

AREAS UNDER THE NORMAL CURVE

Perhaps the greatest usefulness of the normal curve lies in the relationship of the amount of area under the curve lying between certain limits on the base line. In terms of mental-test scores, for example, this simply means the number or percentage of the cases to be expected between two score points. This is because the area under the curve represents the number or percentage of cases. The total area is equal to N , the *total number* of cases. But if we think in terms of a standard curve where $N = 100$, we can readily deal with percentages. For example, 50 per cent of the surface lies above the mean and 50 per cent below. We can also think in terms of a standard curve whose total surface is equal to 1, or unity. In this instance we deal with proportions. The proportion of the area, or cases, lying above the mean is .5 and the proportion below is .5. The statistical tables are given in terms of a total area of 1, and the areas of certain segments are listed as proportions, but it is just as easy to talk in terms of percentages. A percentage is a proportion multiplied by 100, and a pro-

portion is a percentage divided by 100. Thus .46 of the surface is 46 per cent; and 72 per cent of the cases is .72 of the surface, etc.

Proportion of the Area between the Mean and Some Measurement or Score.—We have already had occasion to say that the interval extending one standard deviation on either side of the mean includes about two-thirds of the cases. To say the same thing in another way, from the mean to plus 1σ are to be expected about one-third of the cases, and from the mean to minus 1σ , another one-third of the cases. We can verify this by referring to Table B and looking up the proportion of the

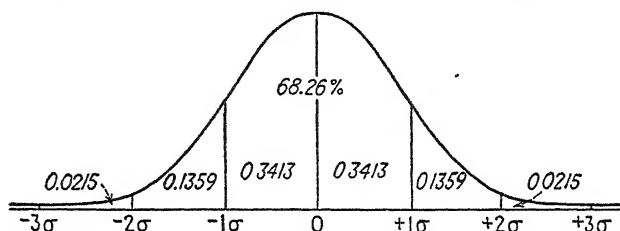


FIG. 14.—Different percentages of area under the normal curve within the various 1-sigma units on the base line.

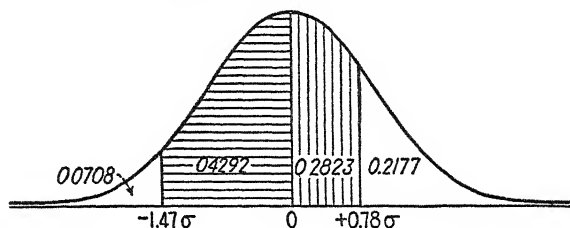


FIG. 15.—Proportions of the total area under the normal curve within certain standard-score limits on the base line

area between the mean and 1σ (*i.e.*, a z equal to 1.00). The area given to four decimal places is .3413, or three thousand four hundred thirteen ten-thousandths of the area. If there were a normal distribution with 10,000 cases, 3,413 of them would be expected between the mean and 1σ . In terms of percentage, it would be 34.13 per cent, or 34.13 cases in 100. The total interval from $+1\sigma$ to -1σ contains twice this area, or .6826, or 68.26 per cent. Figure 14 illustrates these facts graphically. We now see that this is a little more than two-thirds (which would be 66.67 per cent), but with small deviations from normality occurring on every hand we can afford to be so rough with our expectations as to give it as two-thirds.

From Table B, we can also see that between the mean and a point 2σ distant (either above or below, *i.e.*, either $+2\sigma$ or -2σ), we should expect .4772 of the total surface, or 47.72 per cent of the cases.

Included in the range from -2σ to $+2\sigma$, we should find twice this proportion, or .9544 of the area, or 95.44 per cent of the cases. Out to 3σ from the mean extends .4987 of the area, and in both directions from the mean to 3σ we find twice this, or .9974 of the area. Only 26 cases in 10,000 ($10,000 - 9,974$), therefore, should be expected *beyond* the range from -3σ to $+3\sigma$ in a large sample.

To take another example of a less special nature, how much of the area under the normal curve will be found between the mean and $+0.78\sigma$? From the table, we find this to be .2823. In still another problem, how many cases lie between the mean and -1.47σ ? From the table, we find this to be .4292. Figure 15 illustrates these two cases. It will be seen that the positive or negative sign of z merely tells us whether the area extends above the mean or below. The numerical *size* of z , whether positive or negative, determines the *amount* of area between the mean and the point.

So far we have begun each problem of this type with some particular z or standard measurement. Let us start the problem a step or two further back and begin with some raw score or measurement. In the more practical case, we begin with X , not z . In the memory-test data, we may inquire what proportion of the cases come between the mean (26.1) and a score of 35, or a point of 35 on the scale of measurement. This point deviates 8.9 points from the mean ($X - M = +8.9$). This is the deviation x . The standard score z is x/σ , which equals $8.9/6.45 = +1.38$. *Everything must be transformed into standard measure before the probability table may be utilized.* Entering the table with a z of 1.38, we find the corresponding area to be .4162. In other words, 41.62 per cent of the cases in a normal distribution would be found between the mean and 35 points on the scale. In the memory-test data, 41.62 per cent of 86 is 35.8, or, in whole numbers, 36 cases. In a similar manner, which the student should verify, between the mean and a score of 20 are .3276 of the cases, or approximately 28. Between the mean and 15 are about 39 cases of the 86, and if we go on down to a score point of 5, we find 49.95 per cent of the cases.

Special interest attaches to the question of the proportion of cases between the mean and a score of 30.45. It will be found that the standard score corresponding to this is 0.6745. From the table we find that the proportion of the area to this point is .25, or exactly one-fourth. This case is illustrated in Fig. 16. In short, the point at 0.6745σ corresponds to a distance of $1Q$ from the mean.

The Area above or below a Certain Point on the Scale.—For a given deviate or standard score, Table B also gives us the proportion of the

areas above a certain point on the scale or below it. Above a point at $+1\sigma$ will be found .1587 of the area. This is found in column (C) of Table B, because when a vertical line is erected at $+1\sigma$ (see Fig. 17), it divides the total area under the curve into two portions, the one above the line being the smaller of the two. Below the point $+1\sigma$ is the remainder of the area, or the larger portion [found in column (B) of the table], including 8413, or 84 13 per cent of the area. If we were interested in the point -1σ , the larger portion under the curve is now

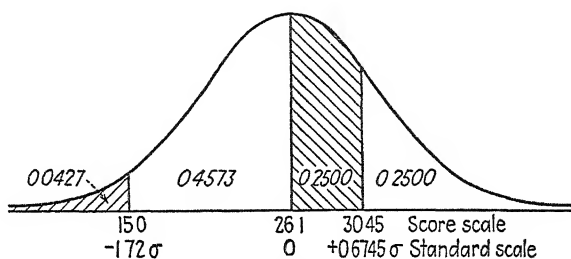


FIG. 16.—Proportions of the area under the normal curve to be expected between certain score limits in the memory test, on the assumption that the distribution is normal

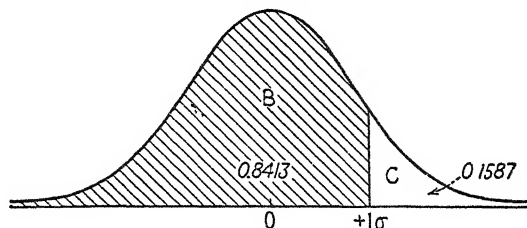


FIG. 17.—Proportions of the area above and below the standard score of $+1\sigma$ and under the normal curve

above the point of division and is found in column (B), whereas the portion below, being the smaller of the two, is found in column (C). The situation is just reversed to the case where the division comes at $+1\sigma$. It is necessary to keep in mind in this kind of problem whether the area we wish to know is under the smaller end of the curve, all on one side of the mean, or whether it is under the larger side of the curve extending across the mean.

The proportion of the area above the point at $+0.78\sigma$ is in the smaller portion, and found in column (C), it is .2177. The area below -1.47σ is also under the smaller portion of the curve, and from column (C), we find that it is .0708 (see Fig. 15). The area above the point -1.47σ would be equal to $1.0 - .0708$, which is .9292. Or it can be found from column (B), since it occupies the larger portion under the

curve, and this also gives us .9292. Or, from Fig. 15, we can see that it is the sum of the area from the point to the mean (.4292) plus .500, which gives the same result

In the memory-test data, where the mean is 26.1 and σ is 6.45, we may ask for the percentage of the cases to be expected below a score of 15. Deviating from the mean 11.1 points, when this is divided by 6.45, we find that the z -score is -1.72 . Corresponding to a z of -1.72 is an area of .0427 in the tail of the normal curve (see Fig. 16). We may expect 4.27 per cent of the cases below a score of 15; or, out of 86, this would be 3.7 cases. Above a score of 15, we should expect the remainder of the cases, naturally; *i.e.*, a proportion of .9573, a percentage of 95.27, and in number of cases, 82.3. Above a score of 30.45, which corresponds to a z -score of $+0.6745$, we should expect 25 per cent of the cases.

Area between Two Points on the Scale.—The first case of this kind of problem has already been mentioned when we asked for the proportion of the area between -1σ and $+1\sigma$ and the like. When the two score points are on two sides of the mean, it is simply a matter of summing the two areas between the mean and the two points. For example, between the points -1.47σ and $+0.78\sigma$, we have the two areas .4292 and .2823 to add (see Fig. 15). The result is .7115, or 71.15 per cent.

When the two points lie on the same side of the mean, it is a matter of subtracting the smaller area from the larger, more inclusive area. For example, the area between points at $+1\sigma$ and $+2\sigma$ can be found by first obtaining from the table the area from the mean to $+1\sigma$ (which is .3413) and the area from the mean to $+2\sigma$ (which is .4772). The area we seek is $.4772 - .3413 = .1359$ (see Fig. 14). The area between points -2σ and -3σ would be the area .4987 [from Table B, column (4)] minus .4772 (from the same source). The difference is equal to .0215, which is illustrated in Fig. 14.

The area between two raw-score points again involves the determination of z -scores as the first step. In the memory-test data, between scores 10 and 20, which correspond to z -scores of -2.50 and -0.945 , respectively, the area is the difference between .4938 and .3276, which is .1662, or 16.62 per cent. The areas from the mean to the two z -scores are found as usual in Table B. As one more example from the same data, the proportion of the cases between scores of 30 and 35 is equal to .1888, for the z -scores are $+0.605$ and $+1.38$, respectively, and the area to the mean in the two cases .2274 and .4162. The student should verify these estimates.

Points above or below Which Certain Proportions of the Cases Fall.

The next problems reverse the processes that have just been described. Before, we were given points on the scale of measurement to determine areas, now we are given areas from which to determine points on the scale. For example, above what point in the normal curve does the highest 10 per cent of the cases come? Ten per cent is a proportion of 10. We could now use Table B in reverse, but it is much more convenient to utilize Table C, which gives the proportions in even steps. We are faced with a problem that gives the proportion in the tail of the curve, so we look in the last column for *C*, the smaller area. We find the *z*-score corresponding to it to be 1.2816. This will be with plus sign, since we are talking about the highest 10 per cent (see Fig. 18). Had we asked below what point does the *lowest* 10 per cent fall,

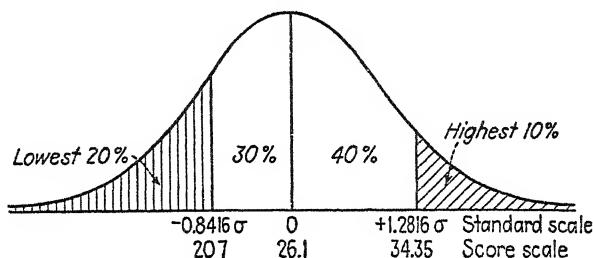


FIG. 18.—Scores above or below which certain percentages of the cases are expected in the memory-test distribution, assuming normality of distribution

the answer would have been -1.2816σ . If the question is, "Above what score lies the highest 80 per cent of the cases?" we are then dealing with the larger proportion under the curve, so we look for the proportion of .80 in the first column of Table C. The corresponding *z*-score is -0.8416σ (see Fig. 18). Had we asked for the point below which is the *lowest* 80 per cent, the answer would have been $+0.8416\sigma$.

To apply these same questions to the memory-test data, we need go a step further and transform the *z*-scores into terms of the raw-score scale. The highest 10 per cent come above a *z* of $+1.2816$. Multiplying this by σ (which is 6.45), we obtain the deviation (*x*) of $+8.27$. The mean (or 26.1) plus 8.27 gives us a score of 34.37 points. The highest 10 per cent in a normal curve with mean of 26.1 and sigma of 6.45 would come above the point 34.37. It happens that this point comes close to the division point between two class intervals, or 34.5. In the actual distribution (see Table 25), 10 cases, or close to 12 per cent, were scores of 35 or above, which is good agreement. Ten per cent would have called for 8.6 cases, or 9 in whole numbers.

The highest 80 per cent of the cases, which we found to come above a z -score of -0.8416σ , will be expected above a raw score of what? The deviation of this point from the mean is -5.43 points, or a score of 20.67 . This comes close to another division point between class intervals, namely, 20.5 . In the actual distribution, 71 or 82.5 per cent of the cases are above a score of 20.5 . Again the agreement between obtained proportion and expected proportion is quite close. To take one more case, which gives a point exactly between class intervals, we ask above what point are 93.2 per cent of the cases? The point turns out to be a score of 16.5 points (the student should verify this). The actual percentage of cases above this score point is 92—again a very close agreement.

Centiles and Corresponding z -scores.—By now it may be apparent that we can look up in the tables the z -score corresponding to any given centile. For example, p_{90} is the point below which are 90 per cent of the cases. Entering Table C with .90 in column (B), we find the corresponding z to be $+1.2816$. Corresponding to p_{80} is the z -score of $+0.8416$. We could find the corresponding raw-score points corresponding to all these z -scores for any particular distribution. If the assumption of normal distribution is valid, this procedure would be an advance step over the recommendation of smoothed ogives for setting up centile norms. But if there is any noticeable skewing in the distribution, this procedure would be rather questionable. The smoothed-ogive method would leave the actual skewness taken into account. Since further measurements with the same test will probably yield the same kind of distribution from the same population, this deviation from normality should be represented in the norms.

It can now be explained how, earlier, (see Table 24, page 72), we arrived at the spacing of centile scores on the profile chart (Fig 10, page 73). The values given to represent the spacing of the centiles are the z -scores corresponding to them, and they were obtained as was explained in the preceding paragraph. The result is to normalize the distribution of all tests, whether the original measuring scale gave a normal distribution or not. There is, in other words, a general underlying assumption of normal distribution of the population in all the abilities represented in the profile chart. The most important gain in so doing is to transform measurements of all abilities into the terms of a common intelligible scale.

The Points between Which Lie Certain Proportions of the Middle Cases.—Among the problems involving area under the curve, there remains the case in which, given the area of a central group, what are

the score limits of that group? The only practical case here occurs when the central group is evenly balanced on either side of the mean, the middle 50 per cent, 80 per cent, or 90 per cent. Those groups, it will be remembered, are significant in connection with indicators of variability and are given distinction in the graphic device illustrated in Fig. 11 (page 74). Here, however, we are talking about the best-fitting normal curve and not the original distribution. The middle 50 per cent extends from Q_1 to Q_3 , or from p_{25} to p_{75} . Going to the tables with a proportion of .75, we find the corresponding z to be, as we should expect, 0.6745 σ . The two points bounding this middle 50 per cent are -0.6745 and $+0.6745$. In the distribution of memory-test scores, these points would correspond to actual scores of 21.75 and 30.45. The interpolated Q_1 and Q_3 in this same obtained distribution were 21.00 and 30.85, respectively, or not very far from those estimated in the best-fitting curve. The middle 80 per cent extends from p_{10} to p_{90} . We have previously determined these to be at a distance of 1.2816 σ , minus and plus. The corresponding raw scores are 17.83 and 34.37. The interpolated 10th and 90th centiles are 17.1 and 35.3, again in close agreement. This kind of problem has really little application in psychological and educational statistics, but is included for the sake of completeness and with the hope that it may lend further insight into the several ramifications of the normal distribution curve. All other problems having to do with area illustrated above do have numerous and valuable applications, some of which we shall meet in the next chapter.

Exercises

1. *a.* Toss six pennies 64 times. After each throw, note and record the number of heads. Compare your obtained frequencies with the expected frequencies. Plot frequency polygons of the two distributions. Compute the mean and standard deviation of the distribution.
- b.* Toss the same six pennies 64 times more, obtaining a new set of data like the first. Compute the mean and standard deviation of this distribution, and make comparisons with the first obtained distribution and with the theoretical distribution.
- c.* Combine the two distributions into a single one. Are the frequencies now any nearer the expected ones? Compute the mean and standard deviation. Are they any nearer the mean and standard deviation of the theoretical distribution?
- d.* One more experiment may be tried in which some of the outcomes with a small number of heads are not counted, but another throw is immediately substituted. Every second case in which at a glance you can tell the number of heads is small, should be ignored and the trial repeated. Again,

obtain 64 record trials. This situation illustrates a biased sampling. What is the effect upon the frequencies?

- e. What would happen in another set of trials if one penny were left head up, only the remaining five being thrown each time but all six coins being observed and all heads being counted?
- 2 Determine the standard scores for all the midpoints in the distribution of Data *J*. Also determine the z -scores for the following raw scores 40, 55, 72, 85, 95
- 3 From Table B, determine the ordinate value at each midpoint of distribution *J*
- 4 Find the expected frequency for each class interval, and tabulate them and the observed frequencies in parallel columns. State some inferences that you can draw from your results

DATA *J*—DISTRIBUTION OF SPELLING-TEST SCORES IN A SUPERIOR GROUP OF FRESHMEN*

Scores	<i>f</i>
82-85	1
78-81	8
74-77	8
70-73	5
66-69	34
62-65	21
58-61	39
54-57	32
50-53	20
46-49	7
42-45	3
38-41	0
34-37	1
Sum	179
Mean	61.1
σ	8.4

* The test was one of the Cooperative series, and the scores are T -scores (see p. 99).

- 5 Find the best-fitting normal curve for Data *J* after the manner of Table 26. Plot the curve along with the obtained frequencies
- 6 Find the proportions and percentages of the areas under the normal curve between the mean and the following z -scores: -2.15 -1.85 -0.19
+0.375 +1.108 +3.52
- 7 Find the proportions and numbers of the cases to be expected between the mean and the following scores in Data *J*: 35 45 60 75 79.5 38.35
- 8 Find the proportions of the area *above* the following z -scores: +2.15 +1.62
+0.175 -0.36 -1.945 -2.875. Also, *below* the following z -scores:
-3.85 -1.225 -0.6745 +0.005 +1.756 +2.385
- 9 Find the proportions and numbers of cases to be expected in Distribution *J* *above* the following scores: 80 55 65 27 69.5 54.5 41.5. And *below* these scores: 85 45 56 35 77.5 41.5 61.5. Whenever possible, compare expected with obtained frequencies

10 Find the proportions of the area falling *between* z -scores of -1.50 to $+1.25$
 -0.05 to $+2.76$ $+0.55$ to $+0.95$ -2.78 to -1.12 $+3.15$ to $+2.95$
 -0.72 to -1.05 $+1.24$ to -0.33

11 Find the proportions and numbers of cases to be expected in Distribution J between scores of 70 and 80 35 and 45 45 and 65 65.5 and 77.5 49.5 and 57.5 45.5 and 65.5 65.5 and 69.6 61.5 and 65.6 53.5 and 57.5
 Whenever possible, compare expected with obtained frequencies

12 Give in terms of standard measurements the points *above* which the following percentages of the cases fall in the normal distribution 85, 55, 35, 42.3, 66.7, and 94.2 per cent

13 Give the z -scores *below* which the following proportions of the cases will fall.
 14 .62 .375 .418 .729

14 *Above* what scores in Distribution J will the following percentages of the cases be expected 12, 54, 84.13, 57.5, and 68.4 per cent?

15 *Below* what scores in Distribution J should we expect the following number of cases 11 63 89.5 123 162? Compare expected with actual cumulative frequencies

16 What z -scores correspond to the following centile points 75 62.5 16.7
 5 99?

17. Between what score limits in Distribution J should we expect the middle 80 per cent of the cases? The middle 50 per cent? The middle 90 per cent? Compare these with the interpolated limits for these same percentages.

CHAPTER VII

SOME APPLICATIONS OF THE NORMAL CURVE

As indicated in the preceding chapter, we shall find here described a number of procedures that depend upon the use of the normal probability curve. They fall roughly into two groups: one kind normalizes distributions and sometimes, in addition, yields general or all-purpose scales of measurement, and the other kind enables us to create measurements out of incompletely metric data, such as rank orders, ratings, and the like.

Normalizing Distributions.—By “normalizing distributions,” we mean to modify scales of measurement in such a way that the resulting frequencies conform to the normal form. It is sometimes believed that the form of distribution of the population from which the sample came is really normal and that the obtained measurements yielded an artificially distorted form of distribution. At other times it is decided that for the sake of convenience the data should be revised in such a manner that the resulting distribution is normal. There are many things to be gained by so doing, as we shall see. Briefly, among them are the advantages of being able to express many different kinds of measurements in terms of a common scale, with the same unit and the same zero point. Most psychological and educational measurements, in their original form, are made in terms of unique measuring scales. A measurement of 45 units or a score of 45 in one instance does not equal a measurement or score of 45 in another. In our search for common ground for the many diverse types of measurement, the normal curve often comes to our rescue. Standard-score scales, *T*-score scales, and other scaled-score systems all depend upon the normal curve as their basis.

Scaling of Observations.—Many quantitative data are obtained, or come to us, as incomplete measurements. Examples of this are judgments in terms of rank order. Several observers, for instance, have placed 25 specimens of art in rank order from best to worst with respect to some artistic quality. Assuming that these 25 specimens do belong at certain positions along a continuum for this quality, but not

at equidistant steps, can we determine those positions and spacings (probably unequal) merely from the judgments of rank order? Other things are judged by observers in terms of a rating scale of some kind or by placing objects in what they suppose are equally spaced groups. If we do not wish to depend upon an observer to keep his imaginary points or groups at equidistant intervals, can we discover and correct for such irregularities? One judge, in using a rating device will mean one thing when he assigns a certain absolute value to an object, and another judge may mean another value when he makes the same response. Is there any way of equating judgments of different observers? Fortunately, if we are willing to make assumptions of normal distributions in the right places, we can answer all these questions in the affirmative and thus take a long step in measuring what was often formerly called the *unmeasurable* and the *intangible*. A few of the procedures for scaling objects when we have only certain human judgments with which to work will be described in an elementary way.

THE USE OF STANDARD SCORES

The Needs for a Common Basis for Comparing Scores.—A student earns scores of 195 in an English examination, 20 in a reading test, 39 in an information test, 139 in a general scholastic-aptitude test, and 41 in a psychological test. Is he therefore best in English and poorest in reading? Could he perhaps be equally good in all the tests? From the raw scores alone, we can answer neither of these questions nor many others that could be legitimately asked. This student's (student I) five scores just cited will be seen listed in column (4) of Table 27. Knowing the means of students in the five tests helps some, for they serve as norms or comparable zero points. The means are listed in column (2). We now see that the student is well above average in English and in scholastic aptitude and is somewhat below average in reading and information, just as the numbers seem to indicate at their face value. The second student, whose raw scores are also in column (4), is numerically highest in the same two and lowest in the same three. When we consider the averages again, however, we find that student II is only about average in English, in scholastic aptitude, and in the psychological test, but he is above average in reading and in the information test.

When a student is above the mean in two tests, in which one is he actually superior? Student I is 39.3 points above the mean in English and 16.2 points above the mean in the psychological test [see column

(5) of Table 27]. Is his superiority in English really greater than his superiority in the psychological test? Student II is 20.3 points above the mean in reading and 17.5 points above the mean in information; is he about equally superior in the two tests?

TABLE 27.—A COMPARISON OF STANDARD SCORES WITH RAW SCORES EARNED BY TWO STUDENTS IN FIVE EXAMINATIONS

(1) Examination	(2) Mean	(3) Stand- ard devi- ation	(4) \bar{X} Raw scores		(5) σ Deviations		(6) z Standard scores		(7) $z - M$ Deviations in standard scores	
			I	II	I	II	I	II	I	II
English	155.7	26.4	195	162	+39.3	+6.3	+1.49	+0.24	.98	.66
Reading	33.7	8.2	20	54	-13.7	+20.3	-1.67	+2.48	2.18	1.58
Information	54.5	9.3	39	72	-15.5	+17.5	-1.67	+1.88	2.18	1.98
Scholastic apti- tude	87.1	25.8	139	84	+51.9	-3.1	+2.01	-0.12	1.50	1.02
Psychological	24.8	6.8	41	25	+16.2	+0.2	+2.38	+0.03	1.87	.87
Sums			434	397			+2.54	+4.51	8.71	5.11
Means							+0.51	+0.90	+1.74	1.02

And how do the two students compare? The superiority of student I as apparent in three tests (English, scholastic aptitude, and psychological) and that of student II, in the other two tests. This we can tell from the raw scores. But suppose the two were competing for a scholarship at a university, which one, if there is to be a choice between the two, should win? The totals of the five scores are 434 and 397, in favor of student I. Granting that the five different abilities are equally important, have we done justice by comparing sums of raw scores? Should we be justified in finding an average of each student's five raw scores?

Suppose that we were interested in determining which student is the more consistent in his abilities, as shown by these five tests, and which one has the greater variability within himself. Would a comparison of the average deviations or standard deviations of the five raw scores give us the answer? As the reader has probably guessed, the reply to most of these questions is in the negative. We are extremely limited in making direct comparisons in terms of raw scores for the reason that raw-score scales are arbitrary and unique. We need a

common scale before such comparisons as we have called for can be made. Standard scores furnish one such common scale

How to Translate Raw Scores into Standard Scores.—As we have previously seen, a standard score z is obtained from a raw score X by means of the formula

$$z = \frac{X - M}{\sigma} \quad (14)$$

In Table 27, we need only to perform the final step with the deviations already given in column (5), dividing by the σ in each case. The standard scores, to two decimal places, are given in column (6).

Now we can answer some of the questions. Student I is most superior in the psychological test, next in scholastic aptitude, and third in English. Had we judged this by his deviations from the mean, we should have decided that his order of superiority was scholastic aptitude first, English second, and psychological third. We find that in terms of standard scores he is equally deficient in reading ability and information, whereas the deviations would have placed him lower in information than in reading. Student II's five standard scores come in about the same rank order as do his deviation scores but certainly not in the same order as his raw scores.

When comparing the two students in terms of raw scores, we should conclude that student I has the greatest advantage in number of points in scholastic aptitude; in terms of deviations, this would be the same, but in terms of standard scores it is in the psychological test that the advantage is greatest. Student II has about the same superiority over student I in the reading and information tests in terms of raw scores and deviations but has decidedly greater superiority in reading ability in terms of standard scores. When we compare the two students as to total or average score, whereas the raw-score total gives student I the distinct advantage of 37 points, or an *average* superiority of about 7 points, the standard-score averages reverse the order and give student II a 0.39σ lead. In a scholarship contest, we should conclude that student II has the greater all-round ability as indicated by these tests, when students are compared on a standard-score basis.

Studies of variability within persons (intravariability) have often resorted to the use of standard scores. In terms of them, is student I more or less variable than student II? Here the average deviation is probably the best mode of comparison. In column (7) of Table 27 is given the absolute deviation of each standard score from the student's own mean score. The average deviations of these two students in the

five tests are 1.74 and 1.02, respectively. In other words, student I is about 70 per cent more variable than student II. Although this is the usual procedure for determining intravariability, a word of caution is important. In using this procedure, we are assuming that in all the abilities measured the true variability of the group measured is the same. The standard-score scale makes the distributions all alike, with standard deviations equal to 1.0. Should we happen to have sampled a group that is actually more variable in one ability than in another, we do not really have comparable units of measurement in all tests. This procedure also assumes tests of approximately equal reliability.

Disadvantages of Standard Scores.—Although standard scores will do for us all that we have said and more, under the proper conditions, there are several things about them which make them less convenient than some others. One shortcoming is the fact that half the scores will be negative in sign, which makes things awkward in computation. Another disadvantage is the very large unit, which is one standard deviation. Although we can give scores to as many decimal places as we should like and thus make use of all the accuracy we could ever justify, decimal numbers are not as convenient as integral scores. As most obtained distributions are, a total range of from five to six units of one standard deviation each is the usual thing for a test scale. If we confined ourselves to integral scores here, our distinctions would be entirely too crude. For these reasons, although the idea of standard scores is basic, we seek other scales with smaller units and all positive scores.

THE *T*-SCALE AND *T*-SCALING OF TESTS

The well-known *T*-scale overcomes the two objections just raised against standard scores and adds besides an advantage peculiar to itself. It adopts as its unit one-tenth of a standard deviation, so that an ordinary distribution with a range of 5 to 6σ on its base line yields 50 to 60 integral *T*-scale scores. In addition, the *T*-scale goes beyond any ordinary distribution, extending its scale over a spread of 10 standard deviations, or 100 units in all. Any age or grade group would yield its own distribution extending 5 to 6σ . A group just higher in ability would overlap this one and yet would need an extension over new units beyond the limit of the first group. A third group of lower age would need an extension of the measuring stick at the other end. When all groups from lowest to highest are taken into account, considerable extension is required. The result, with these extensions, is a

single common scale on which all groups, from highest to lowest, have a common unit and a common zero point. It has been found in practice that a scale with 100 units (or 10σ) will be extensive enough. It is based upon a normal curve whose tails extend from minus 5σ to plus 5σ (see Fig 19). Besides making the unit equal to 0.1σ , the *T*-scale also moves the zero point to the extreme left, placing it at -5σ . The mean now becomes 50, and the other *T*-scale points are distributed as in Fig 19.

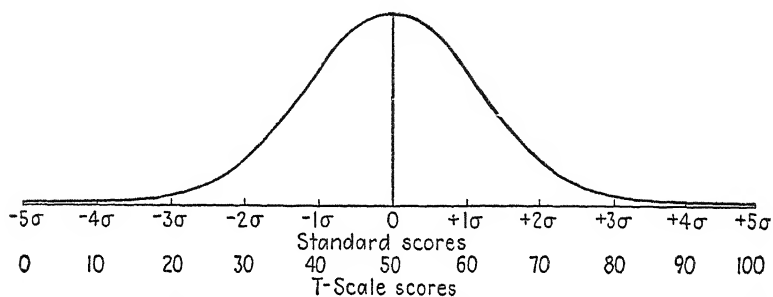


FIG. 19 —The *T*-scale and its relation to the standard-score scale of 10 sigma units

McCall, who originated the *T*-scale, suggested that the mean of this curve should be that of a representative twelve-year-old group. This mean was chosen because the twelve-year-olds are about midway along the scale of mental development. Since any limited sample of them would range over not more than about 60 units of the *T*-scale, groups of higher and lower ability were required to complete the picture and to determine what kind of performance comes at 80 to 100 points at the upper end of the scale and 0 to 20 at the lower end. The method of finding *T*-scale equivalents for performances beyond the ranges of tested samples will not be described here. Suffice it to say that many of the best test makers take pains to set up means of converting raw scores on all tests into *T*-score equivalents. The *T*-scale principle can be used with any standard group of individuals, whether they are twelve-year-olds or not. The procedure for converting raw scores in any test into *T*-scale equivalents (though not with the twelve-year-old mean and unit) will now be described.

How to Set Up *T*-scale Equivalents for Raw Scores.—A college or university or a single school system may wish to use the *T*-scale idea as its common yardstick for all its tests. The freshmen entering a large university, for example, may be taken as the standard group for this purpose. As an illustration, let us use the data in Table 28. Here is a distribution of 83 scores obtained by freshmen in an English

TABLE 28—THE CALCULATION OF T-SCORES FOR A DISTRIBUTION OF ENGLISH-EXAMINATION SCORES

(1) Scores	(2) Upper limit of interval	(3) Frequency	(4) Cumulative frequency	(5) Cumulative proportion	(6) T-score (from Table 29)
225-229	229 5	1	83	1 000	—
220-224	224 5	0	82	988	72 6
215-219	219 5	1	82	988	72 6
210-214	214 5	5	81	976	69 8
205-209	209 5	5	76	916	63 8
200-204	204 5	7	71	855	60 6
195-199	199 5	6	64	771	57 4
190-194	194 5	6	58	700	55 2
185-189	189 5	6	52	627	53 2
180-184	184 5	11	46	554	51 4
175-179	179 5	9	35	422	48 0
170-174	174 5	5	26	313	45 1
165-169	169 5	5	21	253	43 3
160-164	164 5	6	16	193	41 3
155-159	159 5	5	10	120	38 2
150-154	154 5	2	5	060	34 5
145-149	149 5	1	3	036	32 0
140-144	144 5	1	2	024	30 2
135-139	139 5	0	1	012	27 4
130-134	134 5	1	1	012	27 4

examination of the objectively scored type. The procedure will be described step by step:

- Step 1. List the class intervals as usual. Here a maximum number of class intervals is best; 20 or even more.
- Step 2. List the exact upper limits of class intervals.
- Step 3. List the frequencies.
- Step 4. List the cumulative frequencies (see page 64 for instructions).
- Step 5. Find the cumulative proportions for the class intervals.
- Step 6. Find the corresponding *T*-scores from Table 29. These are then listed in the last column of Table 28, given to one decimal place. We usually want finally a ready means of reading directly the *T*-score corresponding to any integral

raw score. It is recommended that the remaining steps be taken to satisfy this objective.

TABLE 29—A TABLE TO AID IN THE CALCULATION OF T-SCORES

Proportion below the Point	<i>T</i> -score	Proportion below the Point	<i>T</i> -score	Proportion below the Point	<i>T</i> -score
0005	17 1	100	37 2	900	62.8
0007	18 1	120	38 3	910	63 4
0010	19 1	140	39 2	920	64 1
0015	20 3	160	40 1	930	64 8
0020	21 2	180	40 8	940	65 5
0025	21 9	200	41 6	950	66 4
0030	22 5	220	42 3	960	67 5
0040	23 5	250	43 3	965	68 1
0050	24 2	300	44 8	970	68 8
0070	25 4	350	46 1	975	69 6
010	26 7	400	47 5	980	70 5
015	28 3	450	48 7	985	71 7
.020	29 5	500	50 0	990	73 3
025	30 4	550	51 3	993	74 6
030	31 2	600	52 5	995	75 8
035	31 9	650	53 9	9960	76 5
040	32 5	700	55 2	9970	77 5
050	33 6	750	56 7	9975	78 1
060	34 5	780	57 7	9980	78 8
070	35.2	800	58 4	9985	79 7
080	35 9	820	59 2	9990	80 9
090	36 6	840	59 9	9993	81 9
		860	60 8	9995	82 9
		880	61 7		

Step 7. Plot a series of points to represent each *T*-score in Table 28 corresponding to the upper limit of the class interval, as in Fig. 20. The points should fall along rather closely to a straight line. The reason that they are not perfectly in line is that there are some irregularities in the original data. Draw through the points with a straight edge a line that will come as close to all the points as seems possible. Among those that do not touch the line, as many of them should

be above it as below it. The line may be extended beyond the ends of the points at both ends.

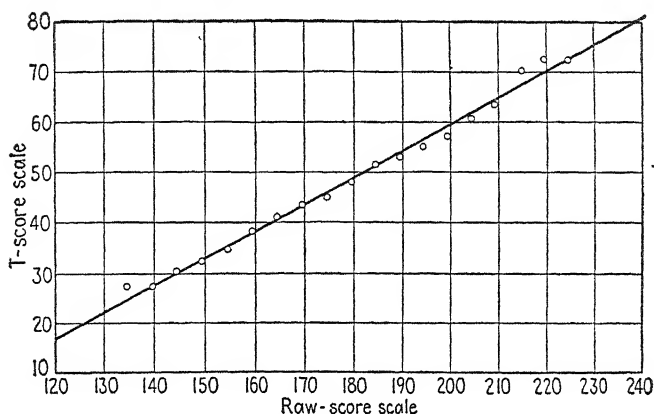


FIG 20—A smoothing process applied in deriving *T*-scale equivalents for English-examination scores (see Table 28)

Step 8. For any integral raw score point, we can now find the corresponding *T*-score points. For example, in Fig. 20, a raw score of 220 corresponds to a *T*-score of 70, and a raw score of 150 corresponds with a *T*-score of 33. In this we favor integral *T*-scores but at times have to resort to half points when we cannot decide upon the nearest unit.

Step 9. Prepare a table in which every integral raw score, or every second, third, or fifth one, appears in one column and the corresponding *T*-scores in the other. Table 30 is such a

TABLE 30—RECTIFIED SCALING WITH *T*-SCORES FOR THE DISTRIBUTION OF ENGLISH-EXAMINATION SCORES

Examination score	<i>T</i> -score	Examination score	<i>T</i> -score	Examination score	<i>T</i> -score
240	81	195	57	155	35 5
235	78	190	54	150	33
230	75 5	185	51 5	145	30
225	73	180	49	140	27 5
220	70	175	46	135	25
215	67 5	170	43 5	130	22
210	65	165	41	125	20 5
205	62	160	38	120	17
200	59 5				

tabulation. It will serve for all future purposes of translation where the original tested group remains the standard.

Some Disadvantages of the *T*-scale System.—The ideal of a single, common, extensive measuring scale for psychological and educational measurements is fine. And there are many places in which it serves well. But in much practical work it has to be supplanted with something less complete. Not every tester can or need take the trouble to work out the necessary *T*-scale equivalents. Furthermore, the unit of the *T*-scale may often be smaller than is justified in view of the imperfect reliability of the tests. *T*-scores then give an appearance and an assurance of discriminations that cannot actually be made. Remember that in many single tests the standard deviation is less than 10 score points, which means that the *T*-scale unit of 0.1σ is then smaller than the original raw-score unit. In the practical work of guidance and clinical treatment in general, too, much rougher distinctions than 0.1σ are all that we require. For these reasons the writer proposes an alternative scale, one whose unit is 0.5σ , and whose practical range is one of 11 units. This will be called the *C*-scale for convenience.

THE *C*-SCALE SYSTEM

The *C*-scale System.—The principles of the *C*-scale and the derivation of *C*-scale equivalents for raw scores are illustrated in Table 31. It is so arranged that the mean will be exactly at point 5.0, with the two limiting classes being called 0 and 10. Column (2) gives the exact limits of the 11 units in terms of standard scores. The corresponding centiles (derived from Table B) are given in column (3). The percentage of cases within each unit is found by subtracting neighboring pairs of centile limits. Thus, in the middle unit, the difference $59.9 - 40.1 = 19.8$, etc. Since it is more convenient to think in terms of whole numbers, the approximate percentages of the cases falling in the different classes are given as nearest whole numbers in column (5). These can be used either as a guide in thinking of the make-up of the standard distribution or even in subdividing lists of scores or individuals when arranged in rank order. Thus, if we had 100 persons lined up in rank order in a test, the highest person would be given the score of 10, the next 3 a score of 9, the next 7 a score of 8, etc., until the last in line is given a score of 0.

But a more exact way of finding raw-score equivalents to *C*-scores is shown in column (6), where we bring the memory-test distribution into the picture. We have already seen how the cumulative percentages or proportions for this distribution were plotted against raw

TABLE 31—THE ELEVEN-POINT SCALED-SCORE SYSTEM AND ITS APPLICATION TO THE MEMORY-TEST DATA

(1)	(2)	(3)	(4)	(5)	(6)	(7)
Scaled score	Standard-score limits	Centile limits	Percentage within each interval	Percentage in whole numbers	Corresponding centiles in the memory-test data	Memory-test scores in each scaled-score interval
10	+2 75	99 7	0 9	1		41+
9	+2 25	98 8	2 8	3	40 5	38-40
8	+1 75	96 0	6 6	7	37 6	35-37
7	+1 25	89 4	12 1	12	34 6	31-34
6	+0 75	77 3	17 4	17	30.8	28-30
5	+0 25	59 9	19 8	20	27 8	25-27
4	-0 25	40 1	17 4	17	24 4	21-24
3	-0 75	22 7	12 1	12	20 8	18-20
2	-1 25	10 6	6 6	7	17 7	15-17
1	-1 75	4 0	2 8	3	14 5	12-14
0	-2 25	1 2	0 9	1	11 8	0-11
	-2 75	0 3				

scores (see Fig. 9). From the smoothed ogive we can read off the raw-score points corresponding to the centile limits given in column (3) of Table 31. These centiles are listed in column (6) of the same table. They divide the raw-score scale up into the 11 units we want. We next decide what integral scores fall within the boundaries set by those limits. The limiting score point between C-scores 9 and 10 is 40.5, which fortunately comes exactly between integral scores of 40 and 41. Raw scores of 41 and up will therefore be in class 10; score 40 will be in class 9. At the next lower limit is the point 37.6. This throws whole score 38 into class 9 and whole score 37 into class 8, and so on down the line, as column (7) will show.

In addition to some advantages already mentioned, it should be pointed out that 11 classes or groups is a number equivalent to the number of class intervals we often choose in grouping raw scores. Future samples in the same tests, whose scores have been transformed into *C*-scores, are already grouped when *C*-scores are assigned. The numbers that stand for those groups are small integral numbers, which makes computations easy. Where punched-card equipment is used, 11 scores require but one column, at the most two, for a test, and the

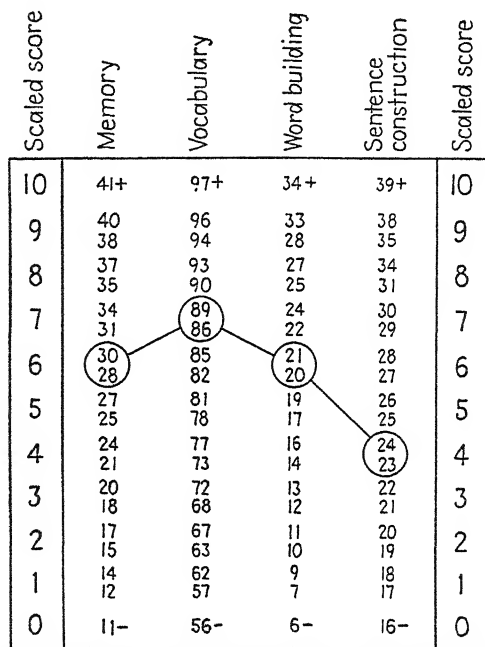


FIG. 21 —A profile chart based upon *C*-scale scores.

coding is already taken care of. The discrimination among individuals, although probably not as fine as the better tests would justify, yet is fine enough both for research work and for clinical purposes. A disadvantage that may be serious is that the *C*-scale has the limited range of 5.5σ , and although this covers the usual practical limits as long as one remains within the sphere of the same kind of individuals, it will naturally cause one to look for supplementation in research work looking beyond these limits.

A graphic profile based upon the *C*-scale idea is presented in Fig. 21. For four tests, which were also used in illustrating the centile profile (see Fig. 10), the limiting scores for every *C*-score unit are written.

This facilitates the locating of any individual's raw scores in the proper *C*-score levels. When the individual's score in a test is located within a certain group, the limiting pair is encircled. The circles are then connected with straight lines, and the profile is complete.

MEASUREMENTS FROM JUDGMENTS OF RANK ORDER

Judgments in terms of rank order are not measurements in themselves, but if we can make certain assumptions about the actual distribution of cases along a genuine metric scale, we can transform rank orders into equivalent measurements. When the things ranked are individuals or the psychological or educational products of individuals and certainly when they are random samples from a normally distributed population, we can assume that the sample approximates a normal distribution. At any rate, it is often safe to assume that individuals or samples near the center of the range are less far apart in quality than those near the extremities. Because this is one of the important characteristics of the normal distribution, except for possible instances of very marked skewing in the sample distribution, the assumption of normality probably does little violence to the situation.

Transforming Ranks into Centile Positions.—From a rank position assigned to any specimen by a judge, we can fortunately specify its centile position, and from its centile position we can determine its equivalent standard-score rating or any other rating based upon standard scores. This is the underlying principle of the most commonly used procedure of deriving measurements from ranks

Usually a rank of 1 means the top individual or specimen in the group, and where there are n things ranked, the lowest item receives a rank of n . The formula for computing the centile position corresponding to a rank is

$$P = 100 \times \frac{n - r + 0.5}{n} \quad (17)$$

where P = centile position.

n = number of things ranked.

r = particular rank assigned to one thing.

In a set of 10 things ranked, for example, a rank of 1 would have the centile position of 95.0, as derived from the formula

$$P = 100 \times \frac{10 - 1 + 0.5}{10} = 100 \times \frac{9.5}{10} = 100 \times 0.95 = 95$$

This is reasonable, when we remember that the top person is conceived as occupying a range that includes the highest tenth of the scale of 100

centile points, or from 90 to 100. The midpoint of this range is 95. The top person in a group of 15 would have a centile position of 96.7, for

$$P = 100 \times \frac{15 - 1 + 0.5}{15} = 100 \times \frac{14.5}{15} = 96.7$$

TABLE 32 —CONVERSION TABLE TO FACILITATE THE TRANSLATION OF RANK ORDERS INTO C-SCALE MEASUREMENTS

Centile-position Range for Each		C-score
C-score Unit		
98 9+		10
96 1-98	8	9
89 5-96	0	8
77.4-89	4	7
60 0-77	3	6
40 2-59	9	5
22 8-40	1	4
10 7-22	7	3
4 1-10	6	2
1 3- 4	0	1
0- 1	2	0

TABLE 33 —DETERMINING T-SCORES AND C-SCORES FROM RANK ORDERS

(1)	(2)	(3)	(4)	(5)
Rank	Number the rank exceeds or is equal to	Centile position or cumulative percentage to the mid-rank	T-score (from Table 29)	C-score (from Table 32)
1	14 5	96 7	68	9
2	13 5	90 0	63	8
3	12 5	83 3	60	7
4	11 5	76 7	57	6
5	10 5	70 0	55	6
6	9 5	63 3	53	6
7	8 5	56 7	52	5
8	7 5	50 0	50	5
9	6 5	43 3	48	5
10	5.5	36 7	47	4
11	4 5	30 0	45	4
12	3 5	23 3	43	4
13	2 5	16 7	40	3
14	1 5	10 0	37	2
15	0 5	3 3	32	1

Similarly, a rank of 2 in 15 has a centile position of 90.0, and a rank of 7 in 15, a position of 56.7. Table 33 gives the solution of all ranks from 1 to 15 in a set of 15 things ranked. Column (2) presents the numerators of the fraction in formula (17) and column (3), the centile positions.

We could now determine the standard scores corresponding to these centile positions, but having previously found fault with the practical use of standard scores we shall here recommend the use either of *T*-scores or *C*-scores. The corresponding *T*-scores are looked up in Table 29, and they are listed in column (4) of Table 33. The *C*-scores corresponding to the centile positions are conveniently looked up in Table 32, especially provided for this purpose. In the opinion of the writer, most rank orders from individual judges are so subject to errors of observation that any scale finer than the *C*-scale for expressing them should be out of the question.

Scaling Ranked Data When Distributions Are Not Normal.—When it is strongly suspected or known that the distribution of cases is not normal, the procedures just described should be seriously modified or replaced. A procedure not assuming anything about the form of distribution of the things ranked has been described elsewhere by the writer.¹

Whenever the form of distribution is fairly well known, certain other procedures are in order. For example, the method of rank order is well adapted to the evaluation of English compositions or themes. Although any single teacher's usual grading of themes is notoriously subjective and faulty, she can place them in rank order for excellence, as judged for certain adopted criteria, which would probably correlate well with another judge's rank order. If it is known that the distribution of scholastic aptitude, or better yet, the distribution of English achievement, is positively skewed, the grading of the themes can then be planned accordingly. The known distribution will tell the teacher what number of A's, B's, C's, etc., should be expected, and having the papers in rank order, she can proceed to assign the expected number of A's to those heading the list, next the expected number of B's, etc. The distribution of marks for the themes will then coincide in mean, variability, and form with the known distribution of the class. This is far better procedure than to adopt the outworn procedure of assuming that every class has a normal distribution and to assign the same

¹ Guilford, J. P., *Psychometric methods*. New York: McGraw-Hill, 1936 Ch. VIII.

percentage of A's, B's, C's, and even F's to every class, regardless of its selection, the kind of teaching received, or the actual achievement.

SCALING JUDGMENTS IN ABSOLUTE CATEGORIES

Persons or things are sometimes judged by being assigned to a small number of defined classes. This is true of rating-scale methods and also of the method of equal-appearing intervals. Although the points on the rating scale and the groups in the method of equal-appearing intervals are presumably equidistant in the minds of the raters, it is often best to treat them as merely *successive* intervals on the scale. In the use of successive intervals or groups is latent the idea of rank order. But in scaling those groups, we do not treat them just as we did ranks. For one reason, there are too few "ranks" in this instance, and for another reason, too many specimens are judged virtually equal by being placed in the same category. In the ranking method, we would have had full rank-order information about these things now seemingly equal within the same category.

Likert's Scaling Procedure.—There are several procedures by which the scaling of things judged in successive intervals can be accomplished. We shall describe one procedure here that has been given prominence by Likert.¹ It assumes a normal distribution of the things rated. As an example, let us take the case in which the common word "recklessness" was judged by 400 students on a scale of five categories in the order: "very unpleasant," "unpleasant," "indifferent," "pleasant," and "very pleasant." The number of students who rated the word in each category may be seen in Table 34. Here it is seen that 58 reacted by marking the word "very unpleasant"; 185 marked it "unpleasant"; 104, "indifferent"; 48, "pleasant"; and 5, "very pleasant." Along the base line of our distribution curve will be placed the so-called *affective scale*, which is a continuum extending from the most unpleasant experience at the left to the most pleasant experience at the right, with a point of absolute indifference at the center. We shall not assume that the five descriptive terms used to describe the five categories are really equidistant psychologically. We do not know what their respective spacings are. We propose to find out by the procedure next to be described.

As usual, we let area under the normal curve stand for frequencies or proportions. Since we are going to work in terms of a curve with unit area, we must deal with proportions. The proportions of judg-

¹ Likert, R., A technique for the measurement of attitudes. *Arch Psychol*, 1932, No. 140.

ments in the categories (frequency divided by N in each case) are given in Table 34. They are illustrated by marking off five divisions under the normal curve in Fig. 22. It will be seen at once that the segments of

TABLE 34—THE CALCULATION OF THE AVERAGE STANDARD MEASUREMENT FOR JUDGMENTS IN ONE OF SUCCESSIVE CATEGORIES
EXAMPLE: DISTRIBUTION OF JUDGMENTS OF 400 STUDENTS ON PLEASANTNESS AND UNPLEASANTNESS OF THE WORD "RECKLESSNESS"

Categories	Very unpleasant	Unpleasant	Indifferent	Pleasant	Very pleasant
Frequencies	58	185	104	48	5
Proportions ($p_2 - p_1$)	0.1450	0.4625	0.2600	0.1200	0.0125
Proportions below the category (p_1)	0.0000	0.1450	.6075	.8675	0.9875
Proportions below, plus those in the category (p_2)	0.1450	0.6075	0.8675	0.9875	1.0000
Ordinate at lower limit of category (y_1)	0.0000	0.2279	0.3844	0.2143	0.0323
Ordinate at upper limit of category (y_2)	0.2279	0.3844	0.2143	0.0323	0.0000
$y_1 - y_2$	-0.2279	-0.1565	+0.1701	+0.1820	+0.0323
z	-1.57	-0.34	+0.65	+1.52	+2.58

the base line, our linear scale, occupied by the respective categories are not by any means equal in width. If we want a single value to stand for each category, the first idea that occurs to us probably is to take

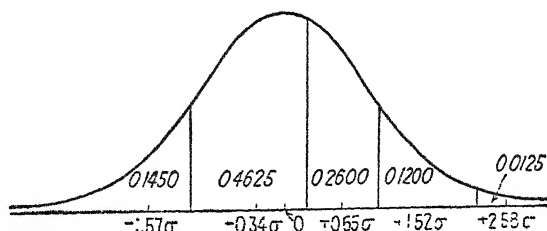


FIG. 22—Segments under the normal curve representing the proportions of judgments in the five categories of pleasantness and unpleasantness. The scale value of each category is the centroid (or arithmetic mean) of the segment

the midpoint of each interval. There are two objections to such a choice in this situation. In the first place, the two end categories stretch off to indefinite limits unless we were to be very arbitrary and

assigned definite outer limits to those end categories. In the second place, the mean of the cases within each group would best represent the cases within the interval, and this will not coincide with the midpoint. We have previously allowed the midpoint of an interval to represent the cases within that interval, but in those instances we had more than 10 intervals as a rule, and they were narrower in range. Here we have only 5. There is always a small error introduced by using the midpoint to stand for all the cases, and the coarser the grouping (the smaller the number of classes) the greater is this error. Here we feel compelled to compute the mean of each group. The mean of any segment under the normal curve between two limits is given by the formula

$$z = \frac{y_1 - y_2}{p_2 - p_1} \quad (18)$$

where z = z -score or standard measurement in terms of which the mean is given.

y_1 = ordinate at the *lower* limit of the segment and is to be determined from Table C.

y_2 = ordinate at the *upper* limit of the segment.

p_1 = proportion of cases *below the segment*

p_2 = proportion of cases *below the upper limit of the segment*.

$p_2 - p_1$ = proportion *within the segment*.

The work of computing the mean of a segment is completely illustrated in Table 34. First are listed the five proportions within the five segments, then the proportions p_1 and p_2 . Next, from Table C, which is entered with p_1 and then p_2 , are found the corresponding ordinates. The next step is to find the differences in ordinate $y_1 - y_2$. Lastly formula (18) can be applied, giving the z values we wanted. The z values are listed in the last row of Table 34 and are shown graphically in Fig. 22.

It can be seen now that the scale separations between successive categories are not equal. From the point for "very unpleasant" to "unpleasant" is a standard distance of 1.23 (*i.e.*, from -1.57 to -0.34). From "unpleasant" to "indifferent" is a distance of 0.99 standard unit. From "indifferent" to "pleasant" the distance is 0.87; from "pleasant" to "very pleasant" it is 1.06 units. The discrepancies could have been worse; but they are serious enough to cause us to hesitate in labeling the original categories with the numbers 1, 2, 3, 4, and 5, as if they were just one unit apart, and computing mean scale positions for words on the basis of this scale.

A Common Scale for All Specimens.—The mean of this normal distribution for the judgments of the word “recklessness” comes at the standard score of zero, and this will be true regardless of what word or other stimulus is being judged. This does not mean that all words have the same average position on the affective scale. On the common scale upon which other words are also to be evaluated, we shall want to anchor the zero point to serve for all cases, and the most reasonable place for this is the indifference point. In this distribution, the “indifference” category came at a point 0.65σ above the mean. If we now shift the zero point up to this position as our point of reference, we shall find that the mean affective position of the word “recklessness” is therefore at -0.65 .

We can similarly find the mean positions of all the words so rated on the same affective scale whose zero point is the point of indifference rating. But there are sometimes obstacles in the way. It is unlikely that distributions for all words will be normal in form. Some may be bimodal, even, and many may be skewed, particularly those words near the ends of the affective scale. Another difficulty is that the real dispersions or variabilities are probably not the same for all words. On some, the judges may agree very closely, and on others, they may differ considerably. The standard-deviation units we have for the word “recklessness” would not necessarily coincide with those for other words. It is better that we try to determine once and for all the relative spacing of the five judgments as determined by several sample distributions and use these facts, assuming that this spacing remains relatively constant no matter what word is being rated.

Following this line of thought and using the word “recklessness” and its distribution as the basis for our scale, the positions of the five categories on this common scale would be 0.65 units lower than those given in the last row of Table 34. Deducting 0.65 from them, we have -2.22 , -0.99 , 0.0 , $+0.87$, and $+1.93$ as the scale positions of the five categories. Should we wish to make the unit of the scale equal 0.1σ , the five values become -22.2 , -9.9 , 0.0 , 8.7 , and 19.3 , respectively. And should we wish integers, rounded, they would become -22 , -10 , 0 , 9 , and 19 . Should we wish to have all positive numbers, we could add to them a constant large enough to make them all positive. If the constant 22 is added, to make the lowest value zero, we have 0 , 12 , 22 , 31 , and 41 . Actually, in this case, the increments are so near to 10 (they are 12, 10, 9, and 10) that one would almost be tempted to forget the discrepancies and revert to a 0, 1, 2, 3, and 4 scale.

But one would not ordinarily permit responses to just one word to determine the spacing of the categories of judgment for all words. In this particular investigation, there were some 400 words rated. It would be wise to evaluate the separations among the five categories at least 20 times, with 20 different words as a basis. Words that yield judgments in all five categories are preferable for this purpose. This would yield at least 20 estimates of the scale separations between the neighboring categories, and a mean of these 20 estimates would give a much more adequate basis for the final scale positions of the five categories. It would be important in this to be sure that the total range of category values is about the same for all words. Since words vary in their dispersions on the affective continuum, the ranges may not be the same. Either adjustments must be made for varying range, or else those words giving ranges differing noticeably from the rest should be left out of account. Space does not permit going more completely into detail as to this procedure.

There are, in addition to the few scaling methods described here, a number of others, but they would take us beyond the scope of this introductory treatment. Here we are confined to the ones most easily applied and the otherwise most practical methods. There are, for example, alternative procedures for scaling judgments in absolute categories. There are also scaling methods for judgments in the form of paired comparisons and other forms of comparative judgments as in the method of first choices.¹

SCALING TEST ITEMS FOR DIFFICULTY

One of the most useful applications of the normal curve is the scaling of test items. A rough idea of the difficulty of an item is gained from the percentage of the individuals who pass or fail it. The greater the percentage of failures the more difficult the item; the greater the percentage of passes the easier the item. But as usual, percentages are expressed by means of *areas* under the normal curve, whereas we are interested in linear distances along a scale. We assume that items *can* be placed along a continuum of difficulty and that each item occupies a characteristic point somewhere on this scale. Figure 23 shows such a scale as the base line of the normal curve, with five items placed at certain standard-score positions along that scale. The zero point on this scale would be an item of median difficulty, an item that can be passed by half and failed by half of a specified group. The scale is relative to the general picture of ability of the group, their median

¹ For a fuller treatment of scaling methods, see Guilford, *op. cit.*, Ch. VII

ability, and their dispersion of ability. On the basis of the group of individuals represented in Fig. 23, a group of freshmen engineers, item 1 had a z -score of -0.71 item 4, a value of -1.37 , and item 55, a value of $+1.44$. Any item with a negative sign is easier than the item of median difficulty for the group, *i.e.*, is passed by more than 50 per cent,

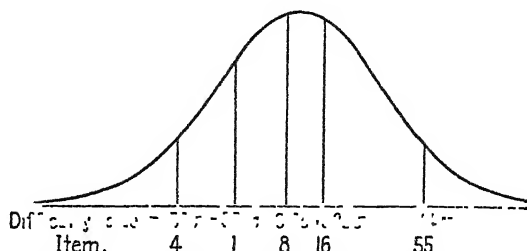


FIG. 23.—The scale values for difficulty of five algebra-examination items and their relation to the normal distribution curve

and any item with a positive sign is more difficult than the item of median difficulty and is passed by less than 50 per cent.

TABLE 35.—DETERMINATION OF ITEM DIFFICULTY OF FIVE ITEMS IN AN ALGEBRA EXAMINATION FROM THE PROPORTIONS OF SUCCESSES IN THREE GROUPS OF STUDENTS

Item number	Relatively homogeneous group of high ability		Relatively homogeneous group of low ability		Large heterogeneous group of students	
	Proportion passing	z	Proportion passing	z	Proportion passing	z
1	95	-1.64	57	-0.18	760	-0.71
4	98	-2.05	85	-1.04	915	-1.37
8	86	-1.08	24	$+0.71$	550	-0.13
16	75	-0.67	05	$+1.64$	400	$+0.25$
55	15	$+1.20$	00		075	$+1.44$

The steps in scaling items are outlined in Table 35. They are as follows:

- Step 1. Determine the proportion of the group passing the item
- Step 2. Look up the corresponding z -score in the tables. If the proportion was greater than .50, give z a negative sign; if less than .50, give it a positive sign.

In Table 35, the same five items are scaled on the basis of the responses of three different groups of students. The first group is

made up of students of high ability in algebra, the second group is made up of low-ranking students, and the third group includes students of widely varying ability. As should be expected, the scale values of the items are generally lowest for the high-scoring group and highest for the low-scoring group. This illustrates the fact that the item values are numerically dependent upon the level of ability of the tested group used as our basis of scaling. It will be noticed, also, that the total *range* of scale values is least for the total group (which has the widest dispersion of ability), and the range is greatest for the two groups of low variability. This is not a defect of the scaling procedure but shows the importance of not comparing difficulty values unless they come from the same population or very similar populations or unless adjustments are made to equate units and zero points for the different groups. This can be done, but we cannot go into the process here.¹ Incidentally, it will be seen that when the proportion is zero (also when it is 1.0), a scale value cannot be estimated, as was true for the low-scoring students with item 55, which no one in this group passed.

Correction for Chance Success.—The items just scaled were from an algebra test in which the student had to find the answer for himself.

TABLE 36—DETERMINATION OF ITEM DIFFICULTY OF NINE ITEMS IN AN ENGLISH EXAMINATION WHEN CHANCE SUCCESS IS A FACTOR

Proportion passing	Number of alternative responses	Corrected proportion of passes	<i>z</i>
.90	4	867	-1 11
.51	4	347	+0 39
.24	4	000	—
.42	3	130	+1 13
.78	3	670	-0 44
.97	3	955	-1 70
.98	2	960	-1 75
.79	2	580	-0 20
.54	2	080	+1 41

Unlike a true-false or a multiple-choice test where the student chooses one out of two or more alternatives, there is almost no chance of success by guessing. In the multiple-choice type of test, however, lucky guessing increases a student's score considerably and also increases the proportion of passing individuals. Proportions that are thus artificially inflated because of the factor of guessing should not be used

¹ See Guilford, *op cit*, Ch. XIII.

TABLE 37.—A TABLE TO FACILITATE THE CORRECTION OF THE PROPORTION OF PASSING INDIVIDUALS FOR A TEST ITEM

p Uncorrected proportion	n Number of alternatives				p Uncorrected proportion	n Number of alternatives			
	2	3	4	5		2	3	4	5
.99	980	985	987	9875	.59	180	385	453	4875
.98	960	970	973	9750	.58	160	370	440	4750
.97	940	955	960	9625	.57	140	355	427	4625
.96	920	940	947	9500	.56	120	340	413	4500
.95	900	925	933	9375	.55	100	325	400	4375
.94	880	910	920	9250	.54	080	310	387	4250
.93	860	895	907	9125	.53	060	295	373	4125
.92	840	880	893	9000	.52	040	280	360	4000
.91	820	865	880	8875	.51	020	265	347	3875
.90	800	850	867	8750	.50	000	250	333	3750
.89	780	835	853	8625	.49	000	235	320	3625
.88	760	820	840	8500	.48	000	220	307	3500
.87	740	805	827	8375	.47		205	293	3375
.86	720	790	813	8250	.46		190	280	3250
.85	700	775	800	8125	.45		175	267	3125
.84	680	760	787	8000	.44		160	253	3000
.83	660	745	773	7875	.43		145	240	2875
.82	640	730	760	7750	.42		130	227	2750
.81	620	715	747	7625	.41		115	213	2625
.80	600	700	733	7500	.40		100	200	2500
.79	580	685	720	7375	.39		085	187	2375
.78	560	670	707	7250	.38		070	173	2250
.77	540	655	693	7125	.37		055	160	2125
.76	520	640	680	7000	.36		040	147	2000
.75	500	625	667	6875	.35		025	133	1875
.74	480	610	653	6750	.34		010	120	1750
.73	460	595	640	6625	.33		000	107	1625
.72	440	480	627	6500	.32		000	093	1500
.71	420	565	613	6375	.31		000	080	1375
.70	400	550	600	6250	.30			067	1250
.69	380	535	587	6125	.29			053	1125
.68	360	520	573	6000	.28			040	1000
.67	340	505	560	5875	.27			027	0875
.66	320	490	547	5750	.26			013	0750
.65	300	475	533	5625	.25			000	0625
.64	280	460	520	5500	.24			000	0500
.63	260	445	507	5375	.23			000	0375
.62	240	430	493	5250	.22				0250
.61	220	415	480	5125	.21				0125
.60	200	400	467	5000	.20				0000

for scaling purposes unless corrections for guessing are made. The correction process is made by means of the formula

$$c p = \frac{n p - 1}{n - 1} \quad (19)$$

where $c p$ = proportion of passes corrected for chance success.

n = number of alternative responses.

p = obtained, uncorrected, proportion of passes.

In Table 36 are given some illustrations of items taken from an English examination. In the first three items, the number of alternative responses was 4, in the next three n is 3 alternatives, and in the last three n is 2. The amount of change in correcting p is seen by comparing column (3) with column (1). In one instance (the third item), the correction yielded a proportion less than zero. Since zero success is the lowest actual p , this could happen only by reason of sampling error, *i.e.*, the item is so difficult that success is purely a matter of guesswork and there was a chance excess of wrong guesses. We call $c p$ zero in this case, and the item is unscalable. A table for correcting proportions of successes is given in the form of Table 37.

TRANSFORMING ONE DISTRIBUTION INTO TERMS OF ANOTHER

In previous pages, we have been concerned with making two or more sets of measurements comparable, giving them a common zero point and a common unit. This was accomplished by translating all of them into a new scale, such as the standard-score scale, the T -scale, or the C -scale. There may be times when we wish to adopt one obtained scale as the common one, letting its mean and standard deviation become the mean and standard deviation for all the others. The procedure for this will be described next.

One practical instance in which this kind of transformation may be useful is in deriving school marks from examination scores. Each examination has its own scale of raw-score points, but there is only one grading system, whether it be the percentage system with a passing point of 60 or 70 or 75 or a letter system or an honor-point system. If it can be decided what the mean and the standard deviation should be for a given class of students in the marking system, then the procedure to be described will enable one to set up rules or equations for transforming raw scores into marks.

Transformation of Ratings.—The illustration of the transformation procedure will be chosen in the sphere of rating-scale evaluations. Assuming that when judge A rates 25 individuals for some particular

trait on an 11-point scale, he maintains somewhat equal units so far as his own judgments are concerned. Assume that judge *B*, in rating the same 25 individuals for the same trait, also maintains equality of unit. But permit *B*'s unit to differ in size from *A*'s, and permit any particular numerical rating, for example, the rating 7, to mean a higher or lower real value for *B* than it does for *A*. These kinds of discrepancy are probably very common among such ratings. When we average ratings obtained from different judges on the same trait, same individual, we are often averaging things quite different in numerical meaning. To be relieved of these constant errors, we should transform the ratings into judgments on a common scale before averaging. We may do so by adopting one judge whose ratings seem to cover the scale

TABLE 38—PARTIAL LISTS OF RATINGS ASSIGNED BY JUDGES *A* AND *B* TO THE SAME INDIVIDUALS FOR THE SAME TRAIT

(1)	(2)	(3)
X_A Ratings by judge <i>A</i>	X_B Ratings by judge <i>B</i>	X_{BA} Ratings by judge <i>B</i> in terms of the mean and sigma of judge <i>A</i>
5	6	3.9
3	7	5.1
2	4	1.5
3	6	3.9
5	8	6.4
7	9	7.6
8	8	6.4
1	4	1.5
7	5	2.7
6	9	7.6
Mean 4.08	6.12	4.07
σ 2.06	1.70	2.08

fairly well as the standard rater and let his mean and standard deviation become the reference values for all distributions.

In Table 38 are given ratings assigned to 10 individuals by judge *A* and also by judge *B*. From a much larger sample of ratings by these two judges, we know that *A*'s average rating is 4.08 and that *B*'s average rating is 6.12. *B* consistently overrates, apparently, as compared with *A*. The standard deviation of *A*'s ratings is 2.06 and of *B*'s, 1.70. *B*'s ratings, taken as a whole, cover less range than *A*'s. This may

mean that *B* has not such great discriminating ability as *A*; or it may mean that he knows people at large who vary more widely than those *A* knows or that he is more cautious, or there may be other reasons. At any rate, we believe that the individuals rated are really just as variable when *B* rates them as when *A* rates them, and they are no higher in the traits rated when *B* rates them than when *A* rates them. We really should assume also that *A*'s and *B*'s ratings are equally reliable, or nearly so. We shall now find what *B*'s ratings should be if he had the same general mean and standard deviation as *A*.

The steps for this procedure are illustrated in Table 39.

TABLE 39—TRANSLATING RATINGS BY JUDGE *B* INTO TERMS OF THE MEAN AND STANDARD DEVIATION OF JUDGE *A*

(1)	(2)	(3)	(4)	(5)
X_B Original ratings by judge <i>B</i>	x_B Deviation of <i>B</i> 's ratings from <i>B</i> 's mean	z Standard meas- urement of <i>B</i> 's ratings	x_{BA} <i>B</i> 's ratings on <i>A</i> 's scale in deviation form	X_{BA} <i>B</i> 's ratings in terms of <i>A</i> 's mean and σ
9	+2 88	+1 70	+3 50	7 6
8	+1 88	+1 12	+2 30	6 4
7	+0 88	+0 52	+1 07	5 1
6	-0 12	-0 07	-0 14	3 9
5	-1 12	-0 66	-1 36	2 7
4	-2 12	-1 25	-2 57	1 5
3	-3 12	-1 84	-3 78	0 3

- Step 1. List the ratings used by judge *B* [column (1)].
- Step 2. Find the deviation of each rating from *B*'s mean rating [column (2)].
- Step 3. Find the corresponding *z*-measurements [column (3)].
- Step 4. Using *A*'s standard deviation (2 06), multiply every *z*-measurement in column (3) by it. This gives deviations from the mean in terms of *A*'s standard deviation. These are in column (4).
- Step 5. Add to each deviation the mean of *A*'s ratings (4 08).

We then have the transformed ratings [column (5)]. Now it is seen that when *B* rates an individual 9 on his scale, he means the same as when *A* rates a person 7.6; when *B* rates a person 7, *A* would probably rate him 5.1 (provided, of course, that *A* gave fractional ratings), etc. On the basis of these transformations, *B*'s ratings of the 10 persons in Table 38 have been changed, as will be seen in column (3). The

mean of this small group of 10 of *B*'s ratings is 4.07, and their sigma is 2.08, both of which are very close to the corresponding values for *A*'s original distribution. In the long run, we should expect them to be exactly the same

A Transformation Equation.—The transformation procedure just given is longer than need be in practice. Its full length was given for the sake of explaining what actually is going on. It is much more expedient to find a simple equation of transformation in the following manner. In general terms, when the measurements in distribution *B* are to be translated into the scale of distribution *A*, the equation is

$$X_{BA} = \left(\frac{\sigma_A}{\sigma_B} \right) X_B - \left[\left(\frac{\sigma_A}{\sigma_B} \right) M_B - M_A \right] \quad (20)$$

where X_{BA} = measurement in distribution *B* transformed into the terms of distribution *A*.

X_B = original measurement in distribution *B*

σ_A = standard deviation of distribution *A*.

σ_B = standard deviation of distribution *B*.

M_B = mean of distribution *B*.

M_A = mean of distribution *A*.

The ratio σ_A/σ_B appears two times in the equation; so we begin by computing it. In our present problem, the ratio is 2.06/1.70, which equals 1.2118. The two means are known, and when substituted in the formula, we have

$$\begin{aligned} X_{BA} &= 1.2118X_B - [(1.2118)(6.12) - 4.08] \\ &= 1.2118X_B - (7.42 - 4.08) \\ &= 1.2118X_B - 3.34 \end{aligned}$$

TABLE 40 —TRANSLATING RATINGS BY JUDGE *B* INTO TERMS OF THE MEAN AND STANDARD DEVIATION OF JUDGE *A* BY MEANS OF AN EQUATION

The equation $X_{BA} = 1.2118X_B - 3.34$

X_B	$1.2118X_B$	X_{BA}
9	10.91	7.6
8	9.69	6.4
7	8.48	5.1
6	7.47	3.9
5	6.06	2.7
4	4.85	1.5
3	3.64	0.3

All we need to do now is to substitute each rating B uses in turn in this equation. Tabulated, this work appears in Table 40. The second column gives the product $12118X_B$ in each case, and the last column comes after deduction of 3.34 in each case. These values coincide with those found by the longer procedure in Table 39.

Exercises

DATA K—MEANS AND STANDARD DEVIATIONS IN FIVE PARTS OF AN ENGINEERING-APTITUDE EXAMINATION (ROUNDED TO WHOLE NUMBERS) AND SCORES OF TWO STUDENTS

Test	Figure classification	Cube visualizing	Syllogism	Paper folding	Form perception
Mean	22	15	28	33	26
σ	4	6	8	5	7
Student A	28	26	30	17	35
Student B	15	32	15	32	41

1. Determine the standard scores for the two students, and draw conclusions as to how the two compare both in terms of raw scores and in terms of standard scores. Which is probably the better student? Which is the more variable student?

DATA L.—FREQUENCY DISTRIBUTIONS OF ENGINEERING FRESHMEN IN THREE APTITUDE TESTS

Scores	Cube visualizing	Syllogism	Form perception
45-49		4	
40-44		13	2
35-39		29	16
30-34	1	42	42
25-29	8	45	52
20-24	35	43	55
15-19	58	24	26
10-14	63	6	13
5-9	36		1
0-4	6		
Sums	207	206	207

2. Determine the equivalent T -scores for one or more of the distributions in Data L. Give the T -scores for the following students in the tests for which you have T -score equivalents

Student	Cube	Syllogism	Form
<i>A</i>	25	22	34
<i>B</i>	5	45	12
<i>C</i>	33	42	37
<i>D</i>	11	20	16

3 Determine the *C*-scale equivalents for the three tests of Data *L*

4 Set up a profile chart for the *C*-score scales of the same three tests Draw in the profiles of the four students whose raw scores are given in Problem 2

DATA *M*—SOME RANK ORDERS OF INDIVIDUALS

<i>n</i> Number ranked	<i>r</i> Three ranks at random
10	2, 6, 9
25	5, 16, 24
36	1, 12, 30
50	4, 21, 50

5 Give the corresponding centile positions of the 12 individuals in Data *M* Transform the ranks given in Data *M* into *T*-scale values, also, into *C*-scale values

DATA *N*—DISTRIBUTIONS OF JUDGMENTS OF TWO WORDS BY 400 STUDENTS

Word	Very unpleasant	Unpleasant	Indifferent	Pleasant	Very pleasant
Gorgeous.	1	12	68	215	104
Slang	16	119	234	30	1

6 Determine the scale positions in terms of *z*-measurements for the five categories in the ratings of the two words in Data *N* Give the scale position of each word with reference to the indifference category and in terms of the standard deviations of the two distributions Are the two standard deviations probably equal? Explain Are the five categories equally spaced? Explain

DATA *O*—PROPORTIONS OF TWO GROUPS OF STUDENTS PASSING ITEMS IN A COMPLETION EXAMINATION

Item number	1	2	5	7	14	22	84
Group I	16	71	47	11	81	28	91
Group II	08	89	68	20	.99	63	92

7 Determine the scale values for difficulty of these items for both groups. Which group probably has higher median ability? Explain. Which group is probably more variable? Explain.

DATA *P*—PROPORTIONS OF STUDENTS PASSING ITEMS IN WHICH THE NUMBER OF ALTERNATIVE RESPONSES VARIES

Number of alternatives	4	4	5	5	3	3	2	2	7	10
Proportion of passes	.47	.87	.30	.98	.36	.84	.99	.75	.82	.29

8 Determine the scale difficulty for the items in Data *P*, both without and with correction for guessing. How is the importance of correcting for guessing shown in your results?

9 In Data *Q*, transform the ratings of judges *B* and *C* into the terms of the distribution of judge *A*. Set up equations of transformation. Interpret your results.

DATA *Q*—RATINGS OF 10 FORMAL DESIGNS BY THREE OBSERVERS

Design	Judge <i>A</i>	Judge <i>B</i>	Judge <i>C</i>
1	9	5	10
2	5	1	5
3	2	3	0
4	7	4	5
5	6	2	3
6	7	7	9
7	4	0	2
8	8	3	6
9	2	4	1
10	6	4	7

CHAPTER VIII

THE RELIABILITY AND SIGNIFICANCE OF STATISTICS

In this chapter, we raise the very important question as to how near to the truth are statistical answers such as means, standard deviations, proportions, and the like. As was said before, any measured sample is frequently made to represent a larger population. Our sampling has to be limited for practical reasons, we cannot measure total populations, or at least it is generally inefficient to do so. Yet we often wish to generalize beyond our sample.

When we obtain the mean of a sample measured in some respect, then, how can we know whether this mean is near the truth for the general population? Can we decide whether the *true* mean is identical with the obtained one or, if not, about how far the obtained mean is from the true one?¹ Fortunately, statistical procedures to be described enable us to find satisfying answers to these questions. Although we can never know the true mean, we can decide within what limits it probably lies and so how much margin of error there is. This we can do if we know the *standard error of the mean*.

THE RELIABILITY OF AVERAGES

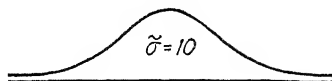
The Distribution of Means of Samples.—Suppose that we are dealing with a large population and we know that it has a standard deviation of 10 units on the measuring scale we are using. Such a distribution is illustrated by the top diagram in Fig. 24. We symbolize this standard deviation by $\tilde{\sigma}$, with the tilde over it to show that it is the standard deviation of the population distribution. It is not necessarily the same numerically as the σ of any sample drawn from this population, though, of course, the latter will be close to the true standard deviation.

The Standard Error (SE) of a Mean.—Suppose, next, that we begin to draw small samples one at a time from the population. To take a reasonable case, let N in each sample equal 25. The obtained,

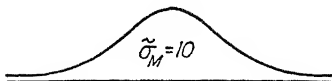
¹ The "true mean" is usually defined as either the mean of the population from which the sample was taken or as the mean of an infinite number of means such as we obtained. So long as we draw samples at random from the same population, the two definitions of true mean will coincide.

or sample, means will not all be the same, any more than all the individuals are the same, but will fluctuate from sample to sample. If we plot a frequency-distribution curve for these means, we find that the distribution is close to normal. In fact, the distribution of a number

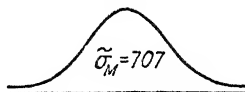
Distribution of individual measures
for a whole population



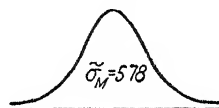
Distribution of means for
samples of one case each



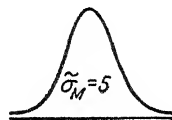
Distribution of means for
samples of two cases each



Distribution of means for
samples of three cases each



Distribution of means for
samples of four cases each



Distribution of means for
samples of 16 cases each



Distribution of means for
samples of 25 cases each



FIG. 24.—Showing the hypothetical decrease in variability or fluctuation of the means of samples as we increase the size of the sample drawn at random from a large population. (Modified from Lindquist, *A First Course in Statistics*, Houghton Mifflin, 1935, p. 155.)

of sample means will in all probability be more nearly normal than will the distribution of individuals within a single sample, even when the population is not normally distributed. The amount of fluctuation or variability among the means is measured by the standard deviation of this distribution to which we give a special name—the *standard error*

of the mean. This value is given by the formula

$$\tilde{\sigma}_M = \frac{\tilde{\sigma}}{\sqrt{N}} \quad (21)$$

where $\tilde{\sigma}_M$ = standard error of the means.

$\tilde{\sigma}$ = true standard deviation of the population.

N = number of cases in each sample (not the number of means in the distribution of means).

Sample Size and the SE of a Mean.—The standard error of the mean is therefore *directly* proportional to the standard deviation of the population and *inversely* proportional to the size of the sample. As the individuals of a population scatter more widely, so will the means of samples drawn from that population also scatter more widely. But as we include more individuals in each sample drawn, the *less* widely can the means scatter about the true mean. In the limiting case, if the sample includes the entire population, the deviation of the sample mean from the population mean can then be only zero, and $\tilde{\sigma}_M$ is zero. In Fig. 24 are shown graphically several instances of samples when N varies. The smallest possible sample occurs when $N = 1$. The mean of each sample is then identical with the individual's measurement in that sample. The dispersion of such means is as great as the dispersion of the total population; $\tilde{\sigma}_M$ then equals $\tilde{\sigma}$, which we have assumed to equal 10. When each sample contains two cases, the

$$\tilde{\sigma}_M = \frac{10}{\sqrt{2}} = 7.07.$$

When each sample contains four cases, $\tilde{\sigma}_M = 10/\sqrt{4} = 5$; etc. The remaining cases in Fig. 24 should now speak for themselves.

Estimating the SE of a Mean from σ .—So far, all that we have said applies to the dispersion of means to be expected when samples are drawn at random from a large population. The $\tilde{\sigma}$ and $\tilde{\sigma}_M$ involved apply only to this theoretical case. We can never know the true $\tilde{\sigma}$ any more than we can know the true mean. Fortunately, statisticians have found a way for us to estimate $\tilde{\sigma}$ and $\tilde{\sigma}_M$ from the data we have. The population standard deviation is related to the obtained standard deviation in the following manner:

$$\tilde{\sigma} = \sigma \sqrt{\frac{N}{N-1}} \quad (22)$$

This means that we can estimate $\bar{\sigma}$, if we wish to do so, directly from the data by the formula

$$\sigma = \sqrt{\frac{\sum x^2}{N-1}} \quad (23)$$

It will be noticed that the standard deviation of the population is practically the same as for the sample except that we have $N-1$ in the denominator instead of N [see formula (8), page 51]. Having thus estimated the true standard deviation, we could then apply formula (21) to compute σ_M . Since this is an estimated standard error (from sample statistics), we should write the symbol as σ_M . Rarely do we feel called upon to estimate $\bar{\sigma}$, whereas we do often compute σ . It is possible to go directly from σ to σ_M if we apply the formula

$$\sigma_M = \frac{\sigma}{\sqrt{N-1}} \quad (24)$$

The last formula is the practical one and the one to remember. Some workers take the expedient course of assuming that $\sigma = \bar{\sigma}$ and substitute σ for $\bar{\sigma}$ in equation (21). This causes little or no error in the calculation of σ_M when N is about 30 or above, but it is generally a source of error when N is less than 30. The writer can see no reason for not being consistent and using the $\sqrt{N-1}$ denominator for all values of N , even when N is large, except when N is by chance an even square, like 100 or 225. In these instances, the simplicity of the arithmetic gained offsets any trivial error introduced by using 100 instead of 99 and 225 instead of 224 as divisors. However, when $N-1$ itself is an even square of some number or when neither N nor $N-1$ are even squares, then $N-1$ would be preferable.

The Interpretation of the SE of a Mean.—We are now ready to apply the standard-error formula to a concrete instance. To revive an old illustration, the ink-blot data, we find that σ is 10.45, and N is 50. Applying formula (24), $\sigma_M = 10.45/\sqrt{49} = 10.45/7 = 1.49$. The standard error of the mean of the ink-blot scores is 1.49. What we are asking when we compute this standard error is how far from the true mean the sample means like the one we obtained would vary. We do not know what the true mean is, but from the value 1.49, we conclude that means of samples of 50 cases each would not deviate from it more than 1.49 units about two-thirds of the time. The interpretation of a standard error of a mean is in the latter respect like that of a standard deviation of a sample. The range from -1σ to $+1\sigma$ in both cases includes about two-thirds of the cases. In the

sample, we know the mean about which the single cases vary, however, whereas in the distribution of means about the true mean, we do not know the value of the true mean.

What is the good, then, of knowing the standard error of the mean? It is this. If we know that two-thirds of the sample means cannot deviate more than 1.49 units from the true mean, we know that our one sample mean also cannot be so very far from the true mean. We can take the next step in interpretation and say that the chances are 2 to 1 that the true mean does not deviate from the obtained one by more than 1.49 points. More definitely, we may say that it is a 2 to 1 bet that the true mean lies between 28.11 and 31.09 (for the sample mean was 29.60, and these are the scale points 1σ below and 1σ above it).

The odds of 2 to 1 are not heavy odds in statistics, though they might be considered so in gambling. If we allowed wider margins of 2σ either way, we could say that the odds are now 21 to 1, that the true mean is within the limits of 29.60 minus 2.98 and 29.60 plus 2.98, or between 26.62 and 32.58. And there are only 27 chances in 10,000 (odds 1 to 369) that the true mean is below 25.13 or above 34.07. The student should review the discussion on area under the normal curve in Ch. VI if the previous remarks are not clear.

Means of Small Samples.—The preceding remarks apply, strictly speaking, when N is very large. When N is small, we should observe the tests of reliability developed during recent years for small samples by Fisher. Instead of speaking of the amount of fluctuation in sample means about the true mean in terms of definite ranges of 1σ , 2σ , or 3σ , there is a growing tendency to adopt instead certain constant odds for the mean being within certain limits.

Fiducial Limits.—To be sure, we spoke of the limits of 1σ giving odds 2 to 1 and the limits of 2σ and 3σ giving odds of 21 to 1 and about 369 to 1, respectively. But in the first place, these odds are not regarded as practically satisfactory, and in the second place, they hold only for large samples. Odds of 2 to 1 are entirely too uncertain for any scientist to bother with, and odds of 369 to 1 are usually unnecessarily one-sided for ordinary purposes. For this reason, Fisher has suggested that we adopt limits (which he calls *fiducial limits*) that include the middle 95 per cent of the values in the one case and the middle 99 per cent in the other. In the first instance the odds would be 95 to 5, or 19 to 1, and in the latter case, 99 to 1. As applied to any particular sampling of means, we are interested in the limits on the measuring scale within which 95 per cent of the sample means would fall or within which 99 per cent of them would fall. In the *normal*

distribution, when N is large, these limits come at 1.96σ from the mean in the one case and 2.6σ from the mean in the other.

Student's t Ratio.—Although the reader recognizes these as standard measures that we have previously denoted by the letter z , for the purpose of testing reliability of statistics, Fisher uses instead the letter t , which was first introduced by a statistician calling himself "Student." Any sample deviating as much as 1.96σ from the true mean in either direction would, in Fisher's terminology, be showing a "significant" deviation. A "significant" deviation from the true mean is thus one that occurs only once in 20 times, and a "very significant" deviation is one that occurs once in 100 times.

Student's Distribution.—With small samples, another revision in the use of fiducial limits is necessary. It has been shown that in order to maintain the same odds for significance, t must increase as N becomes smaller. When N is very large (over 500), the t values of 1.96 and 2.576 correspond to the 95 and 99 per cent levels of significance; but when N is 100, the critical t values must be 1.98 and 2.63—not very much larger, but appreciably so. The reason for this is that whereas the normal distribution applies to t ratios when N is large, when N is small, Student's distribution, which is somewhat leptokurtic, applies. When N is 10, the t 's must be 2.26 and 3.25. It is when N falls below 20 that the change in t 's becomes most marked in order to maintain the same probabilities or odds.

Degrees of Freedom.—In Table D (see page 323) are given the "significant" and "very significant" values of t for different *degrees of freedom*, another concept made important in small-sample statistics. The number of degrees of freedom in dealing with the standard error of the mean is $N - 1$, or one less than the number of measurements in the sample. The same will generally be true for other statistics unless otherwise pointed out in later pages.¹ The student should verify the t 's specified above when N is 100 and 10, as well as examine the table of t 's at the two levels of significance for other values of N .

In the ink-blot problem, a significant deviation from the true mean (one that occurs 5 times in 100) would have a t value of 2.01 (degrees of freedom are 49). In terms of score points, this would be 2.01×1.49 , which is 3.0. The odds are 19 to 1 that the true mean lies between 26.6 and 32.6. A very significant deviation from the true mean would be 2.68σ , which corresponds to a deviation of 4.0 score points. The odds are 99 to 1 that the true mean lies between 25.6 and 33.6. Had we

¹ For an excellent discussion of degrees of freedom, see Walker, H. M. *Degrees of freedom*. *J. educ. Psychol.*, 1940, **31**, 253-260.

adopted the critical t values for large samples (1.96 and 2.576) instead of the ones for small samples (2.01 and 2.68), the conclusions would have been a little different but not seriously so.

In cases where N is much smaller, however, the discrepancy would have been more serious. For example, in the pitch-threshold data where $N = 10$ and $\sigma = 2.27$, the σ_M is estimated as 0.76. Here the fiducial limits have t 's of 2.26 and 3.25 (degrees of freedom are 9). A deviation of 1.72 cycles ($2.26 \times .76$) from the true mean would be very significant. The odds are 19 to 1 that in repeated samples of 10 each, the means would stay within 1.72 cycles of the true mean, and the odds are 99 to 1 that they would stay within 2.47 cycles of it. Reversing the statement, we may say that the odds are 19 to 1 that the true mean lies between 11.48 and 15.92. The odds are 99 to 1 that the mean lies between 10.73 and 15.67. It can now be seen how, although we cannot know the true mean, we can state within what limits it lies with a certain probability, and we can thus infer how much reliance to place upon the obtained mean.

Means of Future Samples.—Note that this says nothing about the means of future samples and where they will lie with respect to the one we obtained. For all we know, the one we obtained may be the highest or lowest within the total range of sample means. *The dispersion of sample means is always around the true mean, which we do not know; never, except by chance, around the obtained mean.* We should not say much by way of prediction of where future sample means will lie, though, of course, they will be expected somewhere near or within the region marked off by the fiducial limits.

Some Words of Caution.—The remarks that precede concerning the standard error of a mean apply only when there has been a *random* sampling from a large population. When there are restrictions imposed upon the process of selecting cases, such as matching groups on the basis of some predetermined quality like mental age, social status, or *IQ*, the fluctuation of means will probably not be so great, and so the *SE* will in reality be smaller than the given formulas show. Alterations in the formulas are then in order. We cannot go into detail with respect to special cases here but refer the reader to a more advanced discussion by Peters and Van Voorhis¹. The formula recommended above is the one almost universally employed, and since more special ones yield a smaller *SE*, we can say that the general formula gives us the most conservative or least reliable picture of the mean. The

¹ Peters, C. C., and Van Voorhis, W. R. *Statistical procedures and their mathematical bases*. New York: McGraw-Hill, 1940. Pp. 132-135.

margin of error may be less than we estimated, but it will probably not be more.¹

The Reliability of a Median.—The variability of sample medians about the true median is about 25 per cent greater than the variability of means. This variability can be estimated by the formula

$$\sigma_{Mdn} = \frac{1.253\sigma}{\sqrt{N}} \quad (25)$$

in which σ_{Mdn} stands for the standard error of a median. As applied to the ink-blot data

$$\sigma_{Mdn} = \frac{(1.253)(10.45)}{\sqrt{50}} = \frac{13.09385}{7.071} = 1.85$$

Two-thirds of the sample medians when N equals 50 will be expected within 1.85 points of the true median. The odds are 2 to 1 that the true median is between 26.4 and 30.1. The t for the range of significant variations of sample medians when N is 50 is 2.01 (see Table D). This gives a score margin of plus or minus 3.72, and the odds are 19 to 1 that the true median lies between 24.5 and 32.0. In a similar manner, the odds are 99 to 1 that the true median is between 23.3 and 33.2, or within a range of plus or minus 4.95 points. It should be said that this estimate of the σ_{Mdn} is valid only when the distribution of medians is normal.

THE RELIABILITY OF OTHER STATISTICS

The Standard Error of a Standard Deviation.—The standard deviation will also fluctuate from sample to sample, and the sample standard deviations will form a normal distribution about the true standard deviation as their mean. This standard error is given by the formula

$$\sigma_{\sigma} = \frac{\sigma}{\sqrt{2N}} \quad (26)$$

Applied to the ink-blot data

$$\sigma_{\sigma} = \frac{10.45}{\sqrt{100}} = 1.045$$

We can now say that the odds are 2 to 1 that the samples of σ will not deviate more than 1.045 points from the true σ . For the fiducial limits

¹ For an excellent discussion of sampling, see McNemar, Q. Sampling in psychological research. *Psychol. Bull.*, 1940, **37**, 331-365.

of 95 and 99 per cent, when t is 2.01 and 2.68, respectively, the margin of fluctuation is ± 2.10 points in the first case and ± 2.80 in the second. We may be fairly certain that the true σ is between 7.6 and 13.2, the odds being 99 to 1 in favor of this inference.

The Standard Error of Q .—The standard error of the semi-interquartile range is given by the formula

$$\sigma_Q = \frac{1.166Q}{\sqrt{N}} \quad (27)$$

In the same illustrative data

$$\sigma_Q = \frac{(1.166)(7.5)}{7.071} = 1.24$$

The interpretation of this standard error is very comparable with that for the other statistics thus far, *i.e.*, in terms of degree of confidence that the true Q lies within certain limits or that any sample Q could deviate by chance a certain distance from the true Q .

The Reliability of Proportions.—Much of our data we find in the form of proportions, or they yield proportions by computation instead of averages. How stable will proportions be in successive sampling? Again we assume that for any phenomenon there is a true proportion for the large population and that any obtained proportion is merely a sample. Whereas in determining the standard error of a mean we did not have to let the obtained mean enter into the equation and we did not even need to know it, in the case of proportions, the true proportion should enter into the formula. The equation is

$$\sigma_p = \sqrt{\frac{pq}{N}} \quad (28)$$

where σ_p = standard error of a proportion (not to be confused with σ_P , the *SE* of a centile).

p = proportion.

$q = 1 - p$.

In order to make any kind of an estimate at all of σ_p , we have to assume that the obtained p is the most probable true p . Since the total outcome of the solution of formula (28) depends relatively more upon the size of N than upon the size of p (except where p is less than .1 or greater than .9) and since the obtained p is somewhere near the true p , we feel justified in making this assumption. Compensating for the greater influence of p upon the estimate of σ_p when p is very small or

very large, when the true p approaches 0 or 1.0, the sampling errors become very much smaller; and so the influence of our discrepancy between the true p and the sample p will be kept very low.

It should be added that the distribution of samples of p 's is approximately normal only when N is reasonably large (at least 20) and when p is greater than .1 and less than .9.

The next chapter will provide a more satisfactory way of dealing with the probable fluctuation of proportions under certain circumstances. The reservations stated here with regard to the SE of a proportion also apply to the SE of a percentage or of a frequency, a discussion of which soon follows.

In a group of 30 students, 18 passed a certain item. We may ask how nearly this proportion (18 to 30, or .6) represents the true status of the population with respect to this item. We assume that the true proportion is .6 in order to estimate the σ_p . N is 30, and q is .4; so the $\sigma_p = \sqrt{(.6)(.4)/30} = \sqrt{.008} = .09$. We may now infer that the odds are 2 to 1 that the true proportion lies between .51 and .69. From Table D, we find that when $N - 1 = 29$, the t 's for the customary fiducial limits are 2.04 and 2.75, which yield margins of plus or minus .18 and .25, respectively. We would be almost sure (99 chances in 100) that the true proportion lies between .35 and .85, which, to be sure, are rather wide limits. Looking at the matter in another manner, our obtained proportion could have arisen from an actual situation in which all the way from 35 to 85 per cent really could pass the item. We are thus not very sure about the real status of the population.

The Standard Error of a Percentage.—If we wished to work in terms of percentages, as the last statement implies, we could do so, remembering that a percentage is 100 times a proportion. The standard error of a percentage $\sigma_{\text{per cent}}$ is

$$\sigma_{\text{per cent}} = 100 \sqrt{\frac{pq}{N}} = \sqrt{\frac{m(100 - m)}{N}} \quad (29)$$

where m = percentage of which we wish to find the SE and the other symbols are as previously explained.

The Standard Error of a Frequency.—A frequency, or the number of cases in a certain category, is equal to N times p the proportion; so the σ_f is equal to N times σ_p , and we have

$$\sigma_f = N \sqrt{\frac{pq}{N}} = \sqrt{Npq} \quad (30)$$

The standard error of the number of cases passing the item referred to above is equal to $\sqrt{30 \times .6 \times .4} = \sqrt{7.20} = 2.7$. Since there were 18 in the passing category, we may say that the odds are 2 to 1 that the true number who can pass it lies between 15.3 and 20.7. With the most stringent fiducial limits, with odds 99 to 1, the t is 2.75, and so the margin is ± 7.4 . The true number of students capable of passing the item is almost certainly between 10.6 and 25.4. This inference and others preceding are valid insofar as the assumption that 18 is the most probable true frequency, or that .6 is the most probable true proportion of passing students, is correct.

It must be remembered that we cannot say much about the fluctuations to be expected in future samples from the same population, for this far we do not carry the assumption that .6 is the true proportion. We see from the preceding work that there is one chance in a hundred that the true p may have been lower than .35 or higher than .85. Had the true p been at one of these extremes, let us say at .35, the margin of fluctuation about *this* mean value could have been about as wide as it was estimated to be around .6. Had .35 been the true p , a fluctuation of .25 would have brought the probability range (odds 99 to 1) to one of .10 to .60 and future samples should then be expected within this range, with our previously obtained one being at the upper limit of it.

THE RELIABILITY OF DIFFERENCES

Of much more practical value than the standard errors of means, proportions, and the like are the standard errors of differences between means and between proportions and the like. In experimental practice, we are perpetually comparing measured results under two conditions that we arbitrarily set up. We ask such questions as whether the eye is more sensitive during stimulation of other sense organs or in the absence of such stimulation; whether boys or girls are more capable in a test of perceptual speed; whether one method of teaching subtraction is superior to another in terms of resulting efficiency. This calls for one set of measurements under the one condition and another set under the other condition and a comparison of means. The statistical question is, "How reliable is the difference between means?"

The SE of a Difference between Means.—Again reliability is indicated by a standard error. The amount of fluctuation in a difference between sample means is naturally related to the amount of fluctuation in the means themselves. The simplest relationship is given by the formula

$$\sigma_{d_M} = \sqrt{\sigma^2_{M_1} + \sigma^2_{M_2}} \quad (31)$$

where $\sigma_{dM} = SE$ of a difference between means.

$\sigma_{M_1} = SE$ of the mean of the first distribution.

$\sigma_{M_2} = SE$ of the mean of the second distribution.

This relationship holds only when the two sets of measurements are independent, i.e., uncorrelated. When we are dealing with matched groups, for example, particularly when individuals are matched pair by pair, the formula will have to be enlarged.¹ But more of that later.

Let us apply formula (31) to a typical problem. A group of 114 men and a group of 175 women were given the same word-building test in which the score is the number of words built out of six letters in 5 min. The results are given in summarized form in Table 41.

TABLE 41—MEANS AND OTHER STATISTICS IN THE COMPARISON OF MEN AND WOMEN IN A WORD-BUILDING TEST

Statistic	Men	Women
N	114	175
M	19 7	21 0
σ	6 08	4 89
σ_M	.572	.371
σ_{dM}	.682	
D_M	1 3	
t	1 91	

The women's mean of 21 0 is 1.3 points higher than that for the men. This mean difference is very small numerically, but in view of the relatively large number of cases in the two samples, we should expect the obtained means to be very close to the true means, and perhaps therefore it indicates a real sex difference. The stability of each mean is indicated by its SE , which is .572 in the case of the men and .371 in the case of the women.

Just as sample means distribute themselves normally about the true mean when N is large, the sample differences between means also distribute themselves normally *about the difference between population means*. We think, therefore, of a distribution of sample differences about some true amount of difference. The σ of this distribution is the σ_{dM} , which in this problem equals $\sqrt{.572^2 + .371^2}$. When solved, this reduces to a σ_{dM} , equal to .682. Interpreted as usual, we may say that the odds are 2 to 1 that the true difference does not exceed 1.3 plus .682, which is 1.982, or fall below 0.618, which is 1.3 minus .682.

¹ See formula (32).

The t Ratio for Differences.—Instead of finding the margin of error for fiducial limits giving 95 per cent significance or 99 per cent significance, as we did in the case of single statistics, it is customary here to find the ratio of the obtained difference to the SE of the difference. This ratio is the obtained t value, and here equals $1.3/.682$, or 1.91 . We next find from Table D the value of t required for the number of degrees of freedom that our data provide. We now have $N_1 - 1$ degrees of freedom for the first set of data, or 113 , and $N_2 - 1$ degrees of freedom for the second set of data, or 174 , because the two sets are independent. We may sum these to obtain the total number of degrees of freedom, which is 287 . From Table D , we find that 287 degrees of freedom is not given, but the quantity is so large that we may take the number 300 and its corresponding t values instead. With this many degrees of freedom, a difference would have to be 1.97 times its standard error to be significant and 2.59 times its standard error to be regarded as very significant. If the difference were at least 2.59 times its standard error, we could say that there is less than one chance in a hundred that it could have occurred by random sampling; and if the difference were at least 1.97 times its SE , we could say that there are less than five chances in a hundred that it could have been that great by sampling. Our obtained difference is 1.91 times its SE ; so it is just below the 95 per cent level of significance and is most certainly below the 99 per cent level. We are not therefore very sure that this much difference did not occur merely by sampling from a population where there is really no sex difference at all; this could happen at least once in 20 samples.

The SE of a Difference in Correlated Data.—When the data are so sampled that there is a correlation between the means in the two variables measured, *i.e.*, so that means in pairs of samples tend to rise or fall together (positive correlation) or tend to be contrasting so that when one rises the other falls (a negative correlation), the SE of a difference is given by the formula

$$\sigma_{d_M} = \sqrt{\sigma^2_{M_1} + \sigma^2_{M_2} - 2r_{12}\sigma_{M_1}\sigma_{M_2}} \quad (32)$$

which is like formula (31) except for the last term, in which r_{12} is the correlation *between the two sets of means*.

Fortunately, under the usual circumstances of random sampling, the correlation between the two sets of means is approximately equal to the correlation between two sets of single measurements in two samples. Since we ordinarily have only two samples with two means from which we could not compute r_{12} , this fact is a great convenience.

But in order to compute the correlation between single measurements, we must have the individual measurements in the two samples paired off two by two in some manner. For example, if the same group of students takes the same word-building test twice instead of two different groups taking it, we have the same individual's score in the first trial

TABLE 42 —STRENGTH OF THE PATELLAR REFLEX UNDER TWO CONDITIONS, TENSED AND RELAXED, FOR 26 MEN, AND DIFFERENCES BETWEEN THEM
(Measurements are in Terms of Degrees of Arc)

<i>T</i> Tensed	<i>R</i> Relaxed	<i>T - R</i> Difference
31	35	- 4
19	14	+ 5
22	19	+ 3
26	29	- 3
36	34	+ 2
30	26	+ 4
29	19	+10
36	37	- 1
33	27	+ 6
34	24	+10
19	14	+ 5
19	19	0
26	30	- 4
15	7	+ 8
18	13	+ 5
30	20	+10
18	1	+17
30	29	+ 1
26	18	+ 8
28	21	+ 7
22	29	- 7
8	4	+ 4
16	11	+ 5
21	23	- 2
35	31	+ 4
26	31	- 5
Σ 653	565	+88
M 25.12	21.73	3.39
σ 7.17	9.25	5.135
σ_M 1.43	1.85	1.03

to pair off with his score in the second trial. Or if, in comparing males and females in the test, we want to standardize our two groups better by taking a brother and a sister from each family or if we pair boy with girl with respect to age, *IQ*, or social status, or all such factors, then if these factors of common family, common age, *IQ*, or social status have any influence on word-building score, they automatically introduce correlation into the two samples. We compute a coefficient of correlation in the manner described in Ch. XI and introduce it into formula (32).

In Table 42, we find two sets of knee-jerk measurements, both from the same 26 men but under two conditions. In the first case (*T*), the subjects were squeezing a hand dynamometer just before the stimulus struck the knee, and in the second case (*R*) the "relaxed" knee jerk was obtained under a relaxed, sitting posture. Will the average man show a real difference in height of knee jerk under the tensed condition, as theory would lead us to expect? The two means, with a difference of 3.39 deg., suggest that the theory is vindicated. But we want to be sure that this large a difference could not have happened by random sampling from a population of measurements in which the true difference is zero.

If we were to assume no correlation between the tensed and normal measurements of knee jerk, we should apply formula (31), or we should apply formula (32) with an r_{12} equal to zero, which is actually the same thing. Such a σ_{d_v} turns out to be 2.34 deg. of arc. The *t* ratio is $3.39/2.34$, or 1.45. This *t* falls decidedly short of the 95 per cent level of significance. We should conclude, erroneously, that although there is some difference in the expected direction, it is not a significant one. So far as these indications go, the true difference could be zero, or even less than zero. If the difference were less than zero here, it would, of course, be in direct opposition to theory.

When we compute a coefficient of correlation between the two sets of measurements, we find it to be $+.83$. This means that the men came rather closely in the same rank order in both the tensed and the relaxed conditions. If a man has a high kick under normal conditions, he will be likely to have a correspondingly high kick during the tensed conditions. If a man is low in the one case, he is likely to be low in the other. That is what the high positive coefficient of correlation tells us. The same kind of situation applies to means of samples. If another group of 26 men had a higher normal average response than this one, it would be likely also to have a higher average tensed response. When means rise and fall together, they tend to maintain the same

difference between them. In the case of a perfect positive correlation ($r = +1.0$), the difference between means would remain exactly constant. If all the sample differences between means were identical, their dispersion would be zero, and, σ_{d_M} would equal zero. We would then be absolutely certain of a true difference in the obtained direction. A correlation of $+ .83$ is less than 1.00, however; so there is still some room for variability among the differences. But from the line of reasoning just completed, we can see that the σ_{d_M} is going to be smaller than it turned out to be when we assumed an r equal to zero.

By the use of the complete formula (32), we find the σ_{d_M} to be 1.03, which is less than half the previous estimate of 2.34. The t ratio is now $3.39/1.03 = 3.29$. A t above 3 is obviously in the "very significant" category. In fact, reference to Table D will show that for 24 degrees of freedom ($N - 2$), which we have here, a very significant t is only 2.80.¹ We therefore feel very confident that there is a real difference in favor of the tensed conditions. This is not saying that we feel sure that the true difference is exactly 3.39, it might be more or less than that. All we are saying is that the odds are 99 to 1 that the true difference lies between 0.51 and 6.27, both of which limits are in the positive direction or in the direction of the theory. The theory is therefore in all reasonable probability correct.

Observations Should Often Be Paired.—In setting up an experiment with two groups of subjects or two groups of measurements for statistical comparison, it is well to pair off cases two by two if at all possible, so that any correlation can be computed. Often when such pairing is not actually carried out, there would still be correlation between means of samples anyway, and the full formula for the SE of a difference cannot then be applied, and the σ_{d_M} by formula (31) is overestimated. It is true that under these circumstances, if the correlation is positive, as is usually the case when there is correlation, we can say that the true σ_{d_M} is smaller and that the correct t ratio is larger than the one we estimated. When we have a significant or very significant t ratio under these circumstances, we can be sure that the t we would obtain by taking into account the positive correlation would be even larger. But one difficulty is that when the t ratio obtained under these circumstances is too small to be significant, we cannot conclude anything in particular. Least of all can we conclude that the true difference is probably zero, for had we considered the correlation, we might have found a significantly large t ratio.

¹ When the two sets of measurements are correlated, the number of degrees of freedom is the number of pairs minus 2.

In pairing off individuals or observations, it is important that the pairing be done on some significant basis. It will not pay to do any pairing except on the basis of some trait that correlates with the measurements on which the two groups are going to be compared. For example, if we were to compare two groups of boys as to ability to do a high jump, one group after training of a certain kind and the control group without such training, it would be important that the two groups be equated as to age, among other things. Ability in the high jump, regardless of training, would be dependent upon age, hence correlated with it. But the ability is probably not correlated significantly with grade earned in arithmetic; so there would be no point in matching the groups on this variable.

The basis upon which to match groups having been decided, there are two common ways of carrying out the matching. One is by pairing cases directly. In the problem just mentioned, for every boy of ten years six months in the one group, one would seek a boy of like age in the other. Small discrepancies may well be permitted at times between pairs. If there are about twice as many cases in the one sample as in the other, matching two boys to one would be the solution. The other common way of matching groups is to ignore individuals as such and simply to try to make sure that the two samples have approximately equal means, standard deviations, and skewness. When this is done and the two variables are correlated, the formula for the standard error is¹

$$\sigma_{d_M} = \sqrt{(\sigma^2_{M_1} + \sigma^2_{M_2})(1 - r^2_{xy})} \quad (33)$$

in which r_{xy} is the correlation between x (the variable on which the groups were matched) and y (the variable on which we are testing the difference). If the groups are matched on the basis of two or more variables, a multiple correlation coefficient is involved (see Ch. XIII).

An SE of a Difference Obtained Directly from Differences.—When individuals have been paired off two at a time, we can find the desired statistics directly from differences between pairs. In Table 42, we find the difference in knee-jerk measurements ($T - R$) given with algebraic signs for every individual. If we sum them and divide by N , we obtain the mean of the differences, which is equal to the difference between the means. If we calculate the *SE* of the mean of these differences, we have σ_{d_M} . The σ_{d_M} is thus obtained in the most direct manner. We do not even need to know the *SE*'s of the two means or the amount of correlation present, yet our direct procedure has taken these

¹ McNemar, Q, *Psychol. Bull.* 1940, **37**, 331-365.

things into account. The σ_{d_M} for the knee-jerk data obtained in this manner is identical with that which we found previously, as it should be. The interpretations and conclusions concerning the mean difference are just the same as usual. This more direct method is very strongly recommended whenever there are any doubts about the applicability of the other formulas.

TABLE 43 — PROPORTIONS OF 400 MEN AND 400 WOMEN WHO JUDGED THE WORDS "TO EXPLORE" AND "SYMPHONY" PLEASANT; DIFFERENCES AND STANDARD ERRORS OF DIFFERENCES, AND t RATIOS

	"to explore"	"sym- phony"	r	Differ- ence	σ_{d_p}	t
Men	8775	6850	342	1925	0234	8 23
Women	8700	8875	395	0175	0180	0 97
Difference	0075	2025				
σ_{d_p}	0235	0281				
t	0 32	7 21				

The Reliability of Differences between Proportions, Frequencies, and Percentages.—Consider the data in Table 43. Here we have the proportions of 400 men and of 400 women students who judged two words as "pleasant" or "very pleasant." The two words were "to explore" and "symphony." Here we can raise two questions concerning each word. Is there any sex difference in the proportion judging the word "pleasant?" And within each sex, is there a significantly greater proportion of "pleasant" judgments for one word than for the other? The differences themselves show that the men favor the word "to explore" slightly more than do the women, the difference in proportion being .0075. The women decidedly more often favor the word "symphony," with an excess of .2025 over the proportion of the men who judge it pleasant. The men find the word "to explore" more pleasing than they do the word "symphony" by a margin of .1925, and the women, on the other hand, find the word "symphony" more to their liking than "to explore" by a small margin of .0175. Which of these differences, if any, are significant or very significant according to the rules we have been following? We can test any or all of them for statistical significance.

The Standard Error of a Difference between Proportions.—The standard error of a difference between two proportions is given by the formula

$$\sigma_{d_p} = \sqrt{\sigma_{p_1}^2 + \sigma_{p_2}^2 - 2r_{12}\sigma_{p_1}\sigma_{p_2}} \quad (34)$$

where σ_{d_p} = *SE* of the difference between two proportions.

σ_{p_1} = *SE* of the first proportion.

σ_{p_2} = *SE* of the second proportion.

r_{12} = correlation of proportions in pairs of samples.

Again, it is fortunate for us that the correlation between proportions is equal to the correlation between single cases. The latter we can estimate from the data. In Table 43, we find that the correlation between men's judgments of the two words is given as +.342 and the correlation for the women is +.395, since both words were judged by the same individuals. But in the comparison between sexes, there was no pairing of individual judgments in any known way, so we may assume that the correlations are zero. On this basis we find the σ_{d_p} between men and women for the word "to explore" to be .0235. The obtained difference of .0075 here yields a *t* ratio of 0.32, which is certainly insignificant. The sex difference on the word "symphony" gives a σ_{d_p} of .0281, which yields a *t* ratio of 7.21. This is so far above the limit for "very significant" deviations that we are very confident about its being true that college women find "symphony" more pleasant than do college men. Men also decidedly prefer "to explore" to "symphony," with the highly significant *t* value of 8.23. Women, however, who find "symphony" more pleasing than "to explore" by an excess of .0175, do not give any sure indication that the true difference is in this direction, for the *t* ratio is only 0.97. The results are somewhat in line with that we should expect, but it can be ventured that some differences that we expected to be true did not prove to be significant and perhaps do not exist at all, for example, where we might have expected a difference between sexes on "to explore," a significant one failed completely to appear. These considerations demonstrate the importance of statistical tests of significance of differences and also the satisfactory kind of conclusions that one is enabled to make after applying those tests.

Differences between Percentages and Frequencies—Similar tests of significance can be made for differences between percentages and frequencies. The uses of percentages and frequencies are here perfectly analogous to the use of proportions as they have been in other connections. An illustration of how to test either of these differences will therefore not be given.

The Reliability of Differences between Standard Deviations.—If we are concerned about differences in variability in two distributions as measured by σ , we can also make statistical tests of significance somewhat like the ones already illustrated. The formula for the

standard error of a difference between σ 's is

$$\sigma_{d\sigma} = \sqrt{\sigma^2_{\sigma_1} + \sigma^2_{\sigma_2} - 2r^2_{12}\sigma_{\sigma_1}\sigma_{\sigma_2}} \quad (35)$$

It is especially to be noted that the r_{12} in this equation, unlike its appearance in others, is squared, for it has been proved that the correlation between standard deviations in pairs of samples is equal to the square of the correlation coefficient between individual pairs of measurements, hence the squaring in formula (35).

We may apply this formula to the data in Table 41 for the word-building test. Here we find the men more variable than the women by a difference of $6.08 - 4.89$, or 1.19 points. Is this difference significant, or could it have arisen as a natural deviation from a true difference of zero, *i.e.*, true equality of the sexes in variability? The $\sigma_{d\sigma}$ proves to be .476 (the correlation being zero) and the t ratio is $1.19/.476$, or 2.50. A very significant t when there are about 300 degrees of freedom is 2.59 (see Table D). The difference of 1.19 points therefore just fails to pass the hurdle of significance at the 99 per cent level. There is just more than one chance in a hundred that if the two sexes are equally variable in this test, such a large discrepancy between their standard deviations could have occurred by sampling.

Concerning Errors of Measurement.—It should be pointed out that we have linked the question of reliability and tests of reliability with deviations of sampling throughout this chapter. In doing so, we have implicitly ignored the question of unreliability of our instruments of measurement themselves. We have been talking as if the instruments were perfectly reliable or free from errors of measurement. This is not the case in psychology and education, for here perfectly reliable measuring instruments are nonexistent. The margins of error that our statistical results show are therefore in part attributable to this kind of unreliability as well as unreliability expected in limited samplings. Any SE of a statistic, like σ_M , σ_p , or σ_d , actually indicates a composite of variations due to limited sampling and to imperfect measurement. Since they do include both factors, when measurements are imperfect, we are making maximum allowance for variations, and when we draw conclusions about the effects of sampling alone, we are erring on the conservative side of the question. Differences, for example, that prove not to be significant when imperfect instruments of measurement are used might prove to be significant if the errors of measurement as such were allowed for.

As yet no one has given this problem more than introductory consideration or has offered fully satisfactory solutions. It is for the

student to keep this qualification in mind when he draws his own conclusions from tests of significance or when he reads those offered by other investigators. A conclusion about the significance or lack of significance of any statistic should end with the phrase *as measured*. Since future measurements of the same kind will probably be of like accuracy, the statement about significance is likely to hold.

A Final Suggestion.—A parting warning is necessary. In studies where biased sampling of groups is involved—and this is true for all studies with matched groups or selected sampling—the way of the unwary investigator is strewn with pitfalls. Whenever vital conclusions are at stake, even the experienced investigator would sometimes do well to seek the best statistical advice on his problem.

ANALYSIS OF VARIANCE

It frequently happens in psychological and educational research that we obtain more than two sets of measurements, each under its own set of conditions, and we want some indication as to whether there are significant differences among the sets. We could, of course, pair off two sets at a time, pairing each one with every other one, and test the reliability of the difference in each pair. The practical difficulty in this approach lies in the number of pairs to be examined when there are, let us say, 5 or more sets. Five sets mean 10 pairs, 6 sets mean 15 pairs; and 10 sets would mean 45 pairs. There is always the possibility that none of the differences would prove significant. What we desire in meeting this situation is some procedure by which we can say in advance whether or not there are *any* significant differences. If the answer to such a preliminary survey is "Yes," we can then examine pairs to see just where greatest significance exists. If the answer is "No," our search is over without further ado.

The new methods of Fisher known as *analysis of variance* are well designed to meet this kind of problem as well as other problems. The real problem here is to determine whether sets of data obtained under varying conditions are sufficiently homogeneous as to be regarded as belonging to the same population. Whether or not we combine distributions into larger composite distributions hinges on the answer to this question. Fisher's test of significance in connection with his analysis of variance is designed precisely to tell us whether sets of data are sufficiently different from one another for us to reject the hypothesis that they arose by random sampling from the same population.

The Meaning of Variance.—The variability in a set of measurements, as is already well known, is indicated by the standard deviation

of the distribution. The term "variance" also pertains to the amount of spread or dispersion of measurements around their mean, but it is measured by the *square* of the standard deviation, in other words, by σ^2 . σ^2 is the mean of the squares of the deviation from the mean or is equal to $\Sigma x^2/N$. A great advantage in using the variance or the squares of the deviations rather than the variability is that variances can be added to or subtracted from one another.

The Variance in a Composite Distribution.—In combining two sets of measurements into a composite distribution, it would not be true that the σ of the composite distribution is merely the sum of the σ 's of the two sets or even an average of the two. The *variance* of such a composite distribution, however, is given by Helson's equation¹

$$\sigma_t^2 = \frac{1}{N} (n_1\sigma_1^2 + n_2\sigma_2^2 + n_1d_1^2 + n_2d_2^2) \quad (36)$$

where σ_t^2 = variance of the composite distribution.

n_1 and n_2 = numbers of cases in the two sets.

σ_1 and σ_2 = standard deviations of the two sets.

d_1 and d_2 = deviations of the means of the two sets from the mean of the composite distribution.

From the equation as it stands, the total variance is a weighted sum of the variances of the two part distributions composing it, plus certain corrections (d_1 and d_2), also squared and weighted, which are necessary because in finding σ_1 and σ_2 , we used deviations from the means of the sets rather than from the mean of the composite. If the means of the sets coincide with the mean of the total, of course, the d 's are zero, and no corrections are necessary. When the d 's equal zero, the two sets are homogeneous in central tendency. In such a situation, there is no more variance among members of the combined sets than there is within the two sets.

Multiplying equation (36) through by N and grouping terms, we have

$$N\sigma_t^2 = (n_1\sigma_1^2 + n_2\sigma_2^2) + (n_1d_1^2 + n_2d_2^2) \quad (37)$$

This puts both sides of the equation in terms of sums of squares of deviations. In the English language rather than mathematical language, it reads that the sum of squares of the deviations of all measurements from the composite mean is equal to the sum of the squares of the deviations *within* the two sets plus the sum of the weighted squares of deviations *between* sets and the composite mean. This is true for

¹ See Guilford, I. P. *Psychometric methods*. New York: McGraw-Hill, 1936. P. 56.

the reason that since

$$\sigma_1 = \sqrt{\frac{\Sigma x_1^2}{n_1}}$$

squaring both sides

$$\sigma_1^2 = \frac{\Sigma x_1^2}{n_1}$$

And so, multiplying both sides by n_1

$$n_1\sigma_1^2 = \Sigma x_1^2$$

$n_1\sigma_1^2$ is the sum of squares of the deviations in set I. In the same manner, it can be shown that $N\sigma_t^2 = \Sigma x^2$, and $n_2\sigma_2^2 = \Sigma x_2^2$. It will be seen, then, that in equation (37) we have effected a clear-cut separation of the contributions (1) of deviations from the set means and (2) of the deviations of set means from the composite mean, to the total variance in the composite. The group $(n_1\sigma_1^2 + n_2\sigma_2^2)$ is the contribution of the variance *within* sets, and the group $(n_1d_1^2 + n_2d_2^2)$ is the contribution of the variance *between* sets.

We have carried on this line of reasoning with the assumption of only two sets in the composite; the same reasoning applies no matter how many sets we are combining. We should merely have to add more terms, one more σ^2 and one more d^2 for each new set included. In general terms, if we let σ_s^2 be the variance of set s , d_s the deviation of the set mean from the grand mean, and n_s the number of cases in the set, the equation can be stated as

$$N\sigma_t^2 = \Sigma n_s\sigma_s^2 + \Sigma n_sd_s^2 \quad (38)$$

The Ratio of "Between" Variance to "Within" Variance.—As was said before, if the d 's are zero, the contribution of the *between* variance to the total variance also is zero. As the d 's become larger, their relative contribution to the total variance becomes greater, and so the total distribution is made up of more heterogeneous elements. The principle of Snedecor's F test is to use the ratio of the *between* variance to the *within* variance as a basis of deciding whether the sets could have arisen by random sampling from the same population. From the sums of squares, which appear in equation (38), we must accordingly compute the two variances.

Degrees of Freedom.—As in all tests of statistical significance, we are dealing with *population* variances, not sample variances. Hence, in order to estimate variances, we should not divide the two expressions on the right-hand side of the equation by the number sampled or N but

by the number of degrees of freedom (see page 130). If we are dealing with k sets of data in each of which are n measurements, the number of degrees of freedom for the *between* variance is $(k - 1)$ and that for the *within* variance is $k(n - 1)$. The two variances are therefore given by the expressions

$$\text{Between variance} = \frac{\sum n_s d_s^2}{k - 1}$$

$$\text{Within variance} = \frac{\sum n_s \sigma_s^2}{k(n - 1)}$$

TABLE 44—WORK SHEET FOR THE ANALYSIS OF VARIANCE IN FOUR SETS OF MEASUREMENTS ON THE GALTON BAR
The Measurements (X)

Set I	Set II	Set III	Set IV	
114	119	112	117	
115	120	116	117	
111	119	116	114	
110	116	115	112	
112	116	112	117	
ΣX_s 562	590	571	577	2,300 ΣX
M_s 112 4	118 0	114 2	115 4	115 0 M
Deviations within Sets (v_s)				
+1 6	+1 0	-2 2	+1 6	
+2 6	+2 0	+1 8	+1 6	
-1 4	+1 0	+1 8	-1 4	
-2 4	-2 0	+0 8	-3 4	
-0 4	-2 0	-2 2	+1 6	
Squares of Deviations within Sets (v_s^2)				
2 56	1 00	4 84	2 56	
6 76	4 00	3 24	2 56	
1 96	1 00	3 24	1 96	
5 76	4 00	0 64	11 56	
0 16	4 00	4 84	2 56	
17 20	14 00	16 80	21 20	69 20 Σv_s^2
Deviations of Set Means from Grand Mean (d)				
d	-2 6	+3 0	-0 8	+0 4
d^2	6 76	9 00	0 64	0 16
nd^2	33 80	45 00	3 20	0 80
				16 56 Σd^2
				82.80 $n \Sigma d^2$

The expression $n_s\sigma_s^2$ is equal to Σx_s^2 , as was said before. We may therefore substitute Σx_s^2 in the last equation. And since in most practical applications of analysis of variance the sets have equal n 's, we may write the two equations

$$\text{Between variance} = \frac{n\Sigma d^2}{k-1} \quad (39)$$

$$\text{Within variance} = \frac{\Sigma x_s^2}{k(n-1)} \quad (40)$$

The subscript may now be dropped from n , since it is a constant throughout, and also from the d 's, but the subscript s is left on the x^2 to indicate that we are here dealing with the deviations from the means of the sets rather than from the grand mean of the composite.

The Solution of an Analysis-of-variance Problem.—In Table 44, we have four sets of observations made by the same individual on the Galton bar. With a constant horizontal line of 115 mm., the subject adjusted another line to seem equal to it. The four sets were obtained under four different arrangements of conditions under which the adjustments were made. Is it likely that the observations all came by random sampling from the same general "population" of adjustments, or were there systematic differences among sets sufficient to say that the data are really not homogeneous? The following steps are followed in the solution of the type in Table 44:

- Step 1. Compute sums and means of the sets, also, the grand total ΣX and the grand mean M
- Step 2. For every set, compute the deviations from the set mean M_s . These are equal to $(X - M_s) = x_s$.
- Step 3. Square the deviations within sets to find each x_s^2 . Sum these to obtain Σx_s^2 , the sum of the squares of deviations within sets.
- Step 4. For each set, compute d , which equals $(M_s - M)$.
- Step 5. Square each d , and find $n\Sigma d^2$.

With these calculations completed (see Table 44), we have the values we need for formulas (39) and (40). The Σx_s^2 is 69.20, and the $n\Sigma d^2$ is 82.80. Dividing these by the appropriate degrees of freedom, we obtain the variances. For this purpose, we set up Table 45. Listing first the degrees of freedom and sums of squared deviations for "between sets" and dividing, we obtain 27.60 as the variance contributed by the d 's. For the corresponding values for "within sets," we find 4.325 as the variance contributed by the x_s 's. The F ratio is 27.6/4.325, which

equals 6.38. The "between" variance is over 6 times as great as the "within" variance.

TABLE 45.—THE TOTAL VARIANCE IN THE GALTON-BAR DATA SUBDIVIDED INTO TWO COMPONENTS

Components	Degrees of freedom	Sums of squares	Variance
Between sets	3	82 80	27 60
Within sets	16	69 20	4 325
Total	19	152 00	

$$F = \frac{27\ 6}{4\ 325} = 6\ 38$$

The significance of an F ratio of this size is determined by reference to Snedecor's table (Table F, page 326). In using this table, we have to consider the two degrees of freedom. For the larger variance, with 3 degrees of freedom, we look for the column in Table F that is headed (3). For the smaller variance, with 16 degrees of freedom, we look down the left-hand margin for the row headed (16). We must interpolate, since row (16) is not given, and thus we find that an F of 3.24 is significant at the 5 per cent level and an F of 5.29 is significant at 1 per cent level; *i.e.*, the odds are 5 to 95 that so large an F as 3.24 could have occurred in a really homogeneous population, and they are 1 to 99 that an F as large as 5.29 could have occurred likewise. Our obtained F is greater than that for the 1 per cent level and so is regarded as very significant. We conclude that there are significant differences among our sets. The test does not tell us where those differences are or whether all of them or only one is significant. To determine this would require further search. We only know from the F test that some significant law of variation between sets does exist. Further examination will be needed to tell us what and where the causes of differences lie.

Computation of Variances from Original Measurements.—Just as we can compute standard deviations, and so variances, from original measurements without computing separate deviations from the means, so we can calculate the necessary constants for an analysis of variance. Such an approach would require us to square the original measurements. With good calculating machines available, this is no large order, but with only pencil and paper it amounts to considerable labor.

Fortunately, by a process of coding, we can bring the numbers down to small size. For the three-place numbers in Table 44, we may sub-

tract the constant of 110 from every one of them, leaving the remainders shown in the first part of Table 46. The variances will not have been

TABLE 46—SOLUTION OF AN ANALYSIS OF VARIANCE FROM ORIGINAL MEASUREMENTS
(Without Determining Deviations from Means)
Measurements (Reduced) (X')

Set I	Set II	Set III	Set IV	
4	9	2	7	
5	10	6	7	
1	9	6	4	
0	6	5	2	
2	6	2	7	
$(\Sigma X')_s$ 12	40	21	27	100 $\Sigma X'$
$(\Sigma X')^2_s$ 144	1,600	441	729	5 M'
				2,914 $\Sigma(\Sigma X')^2_s$
Squared Measurements (X'^2)				
16	81	4	49	
25	100	36	49	
1	81	36	16	
0	36	25	4	
4	36	4	49	
$(\Sigma X'^2)_s$ 46	334	105	167	652 $\Sigma(\Sigma X'^2)_s$

affected in the least by this coding process, for the new values, which we shall call X' , maintain the same distances from one another and from the means, as they did before coding. The sums of squares we need for equations (39) and (40) are found by the following procedure. The sum of the "between" variations squared is given by

$$n\Sigma d^2 = \frac{\Sigma(\Sigma X')^2}{n} - (\Sigma X')(M') \quad (41)$$

The sum of the "within" variations squared cannot be found directly, but since this sum and the one for "between" variations together make up the sum of the squared variations in the total distribution that we *can* compute directly

$$\Sigma x_s^2 = \Sigma x^2 - n\Sigma d^2 \quad (42)$$

and

$$\Sigma x^2 = \Sigma(\Sigma X'^2)_s - (\Sigma X')(M') \quad (43)$$

The steps called for by these formulas are as follows:

- Step 1. Sum the coded measurements X' for each set, to obtain $(\Sigma X')_s$ for each set (see Table 46), and sum these values to obtain $\Sigma X'$. Determine the mean M' to several decimal places.
 - Step 2. Square the sums of the scores to obtain $(\Sigma X')^2_s$ for each set. Accumulate these to find $\Sigma(\Sigma X')^2$.
 - Step 3. Square all the coded measurements to find the X'^2 values.
 - Step 4. Sum all the squared measurements to obtain $\Sigma(\Sigma X')^2$.
- Now, by formula (43)

$$\Sigma x^2 = 652 - (100)(5) = 652 - 500 = 152$$

And by formula (41)

$$n\Sigma d^2 = \frac{2,914}{5} - 500 = 582.8 - 500 = 82.8$$

By formula (42)

$$\Sigma x^2_s = 152.0 - 82.8 = 69.2$$

A comparison of these values with those in Table 44 will show that we have arrived at the very same sums of squares. From here on, the computation of variances and of F ratio is just the same as it was before. The same formulas, (41) and (43), apply also to original measurements without coding.

General Uses of Analysis of Variance.—There is insufficient space here to do more than to give this brief introduction to the analysis-of-variance methods. There are many and varied applications of this simplest case—the separation of variance among a few sets of data into the “within” and “between” variances—both in psychology and in education.

Sets of data may be divided according to chronological-age groups, mental-age groups, sex-difference groups, etc. In psychophysical experiments, judgments of phenomena may be made under various conditions—ascending series versus descending series, variable stimulus first versus second, right versus left, and with many other kinds of variation of conditions, not to speak of individual differences among observers and observations obtained at different times of day or under different states of fatigue or rest or after different degrees of practice. In education, the testing of different teaching methods can be done in different schools, in different classes within the same school, and with different teachers.

It will be recognized that the conditions affecting sets of measurements often vary in more than one direction at the same time. This

complicates the analysis-of-variance solutions in various ways. For further descriptions of how to adapt the method to various kinds of experimental problems, the reader is referred to books that treat the subject at much greater length.¹

By way of hasty evaluation of the method, it may be said that the analysis of variance undoubtedly provides a powerful tool of working through data in order to see where the significant lines of cleavage lie and so in establishing the presence of laws. It can also be said that the method requires the supplementary procedures for a more detailed study of data and that there are other statistical methods—for example, correlation procedures—that enable us to accomplish the same purpose in many instances. Be that as it may, the student will probably see the variance methods brought more and more into use in the solution of educational and psychological problems.

Not the least of its merits is the rather strict set of requirements it presupposes in the designing of experiments. Experimental designs have generally been observed, particularly in psychophysical research, for a long time. But they have generally not been so consciously considered or so well planned so as to yield the maximum number of dependable answers as is true when the experimenter has kept clearly in mind the corresponding statistical tests that go with those designs. The subject of experimental design is well treated in the book by Lindquist already cited, so far as certain educational problems are concerned. Discussions of designs for psychological and other experiments may be found elsewhere.²

Exercises

DATA R—RESULTS FROM A TEST OF THE ABILITY TO NAME FACIAL EXPRESSIONS IN THE RUCKMICK PHOTOGRAPHS

Statistic	Men	Women
<i>N</i>	95	164
<i>M</i>	21 1	22 0
σ	3 62	3 15
<i>Q</i>	2 38	2 16
<i>Mdn</i>	21 5	22 2

¹ See especially Lindquist, E. F., *Statistical analysis in educational research*. New York: Houghton, 1940. Also, Snedecor, G. W., *Statistical methods*. Ames, Iowa: Collegiate, 1937.

² See Baxter, B., *Problems in the planning of psychological experiments*. *Amer J. Psychol.*, 1941, 54, 270-280. Also, Fisher, R. A., *The design of experiments*. Edinburgh: Oliver and Boyd, 1935.

154 FUNDAMENTAL STATISTICS IN PSYCHOLOGY AND EDUCATION

DATA *S*.—QUANTITY WRITTEN IN SENTENCE CONSTRUCTION FROM 10 SETS OF 3
NOUNS EACH AND 10 SETS OF 3 VERBS EACH

Measurement Is the Number of Sentences Written in a Limited Time Subjects
Were 55 Girls

Statistic	Nouns	Verbs
M	24 7	22 8
σ	6 31	5 42
$r_{VV} = 87$		

DATA *T* —NUMBER OF STUDENTS IN TWO GROUPS WHO PASSED EACH OF THREE
ITEMS IN AN INTRODUCTORY PSYCHOLOGY EXAMINATION

	Group I	Group II
N	37	63
Item <i>A</i>	24	26
		$r_{AB} = 19$
Item <i>B</i>	33	32
		$r_{BC} = 32$
Item <i>C</i>	30	44
		$r_{AC} = 25$

1. Compute the standard errors of the means for Data *R*, and interpret your results.
2. Compute the standard errors of the means for Data *S*, and interpret your results
3. Compute the standard errors of the medians for Data *R*, and interpret your results
4. Compute the standard errors of the standard deviations in either Data *R* or Data *S*, and interpret your results
5. Compute the standard errors of the frequencies of passing students in Data *T*, and interpret your results Do the same in terms of percentages and proportions
6. Compute the standard error of the difference in means for Data *R* and also for Data *S*, and test for significance State interpretations
7. Compute the standard error of the difference between medians in Data *R*, and interpret your results
8. Determine the reliability of the differences between standard deviations in Data *R* and *S*. Draw conclusions
9. Determine the reliability of differences between Groups I and II in terms of frequencies, percentages, or proportions of correct responses Interpret your results.
10. Determine the reliability of the differences between proportions passing items *A*, *B*, and *C* for either Group I or Group II. Give your interpretations.

DATA *U* —FIVE SETS OF MEASUREMENTS OF THE LOWER THRESHOLD FOR PITCH

Set I	Set II	Set III	Set IV	Set V
16	18	15	15	17
19	19	15	19	15
17	18	14	16	14
14	17	12	18	16

11 What is the likelihood that all sets of measurements in Data *U* came from the same set of conditions? Apply the *F* test, and discuss your results

CHAPTER IX

TESTING HYPOTHESES

We have already emphasized the point that experiment and statistical method go hand in hand. The one supplements the other. The experiment directs our observations and yields data. By means of statistical methods, we can summarize those data, interpret them, and determine their reliability.

The best experiments are those that are set up to test the truth or falsity of some hypothesis. From previous experience, we believe a certain thing to be true, but it requires a crucial test to enable us to accept or to reject the hypothesis. If the result comes out one way, the hypothesis is probably correct; if it comes out another way, the hypothesis is probably wrong. The term "probably" is inserted because there is no such thing in science as absolute certainty. We are only more or less sure that the result points to one conclusion rather than to another.

The assurance of a conclusion may be of any degree of intensity from "doubtful" to "maybe" to "very likely" to "almost certain" to "practically certain." Statistical procedures give more definite meaning to those degrees of doubt and assurance. In this chapter, particularly, we shall be concerned with giving those concepts more exact meaning, so that we may be able to conclude whether certain outcomes of observations could perchance have arisen by accident or whether they point to something definitely not accidental.

THE NULL HYPOTHESIS

Meaning of the Null Hypothesis.—In recent years, we have been hearing more and more of the expression *null hypothesis*. In very general terms, this hypothesis merely states that in the experimental situation, or even in the nonexperimental situation, whenever things are enumerated or measured, it is assumed for the sake of argument that nothing but the laws of chance are operating. A null hypothesis can be applied in many ways, but an illustration from experiments on extrasensory perception (ESP) is very suitable.

Suppose that an experiment with the Duke University ESP cards is properly set up to prevent the receiver from being influenced by any

cues except possible telepathic stimulation. There are five different symbols on the cards, and in a thoroughly shuffled deck they should come up at random. As each one comes up and an experimenter reads it silently, the receiver makes his judgment. The card is returned to the deck, which is reshuffled, and the next one to be transmitted is selected. Starting with the hypothesis that there are no factors (*including* ESP) at work to determine the receiver's responses, we should expect in the long run an average success of 20 per cent right, or 1 in 5. If any receiver gives an excess of correct responses over and above 20 per cent, we still have to determine whether this excess is significant or whether it could have occurred by the processes of sampling in his limited number of trials. If the excess is one that could have happened as much as once in 10 times (one sample of this size out of 10 such samples), we should still say that the null hypothesis is quite plausible. We could not say that it is certainly established, but we would by no means give it up. Even if the excess over 20 per cent were one that could happen less than once in 20 samples, though we should be more skeptical of the null hypothesis, we should be unjustified in completely rejecting it. When so large a discrepancy as we obtained could occur by sampling less than once in 100 times, we customarily reject the hypothesis. We then say that it is almost certainly not true.

But note that this does not automatically always prove that the alternative hypothesis is true. It does tell us that something other than guesswork is going on, but it does not tell us what that "something other than guesswork" really is. If our experiment is set up so as to exclude all other possible factors than ESP in this case, then, having reduced the crucial experiment to an either-or proposition, *i.e.*, either laws of chance or ESP, and having proved the chance hypothesis wrong, we can accept the ESP hypothesis as true. Unfortunately, the identification and control of all other factors favoring correct responses here is exceedingly difficult. But, in general, the establishment of an experimental fact depends upon it. We shall see shortly how a statistical test of the null hypothesis can be made for this type of experiment; but first let us consider some simpler cases.

Direct Determination of the Probable Validity of a Null Hypothesis
Our first example is a simple psychophysical test situation. A student asserts that he can distinguish between two tones whose stimuli differ only 2 cycles per second. That is his hypothesis: that he possesses genuine power to discriminate this difference in pitch. We doubt him, thus automatically adopting a null hypothesis. Out of 6 trials, how many should we require him to judge correctly before we give up

our hypothesis and yield to his? Our hypothesis implies that when he judges the pair of tones he might just as well flip a coin and report "first higher" for "heads" and "second higher" for "tails." By such guessing, we should expect him to be correct half the time or 3 times out of 6. But how much of an excess over 3 correct judgments will it take to convince us that he is not merely guessing?

In a set of 6 trials, there are 7 possible outcomes—all the way from 6 down to 0 correct judgments. In Table 47 are listed the three cases in which we are interested, 4, 5, and 6 correct judgments. According to the probabilities involved in this case, in 64 samples of 6 judgments each, we should expect only *one* "score" of 6; we should expect a score of 5 right 6 times in 64 and a score of 4, 15 times. The null hypothesis here calls for a score of 6 one-sixty-fourth of the times, a score of 5 six-sixty-fourths of the times, and a score of 4 fifteen-sixty-fourths of the times.

TABLE 47 —EXPECTED OCCURRENCES AND PROBABILITIES OF SPECIFIED NUMBERS OF CORRECT JUDGMENTS IN MAKING SIX JUDGMENTS AT RANDOM

Number of correct judgments	Times expected in 64 sets of judgments	Probability of this number occurring at random	Probability of as many or more judgments occurring at random
6	1	$\frac{1}{64}$	$\frac{1}{64}$
5	6	$\frac{6}{64}$	$\frac{7}{64}$
4	15	$\frac{15}{64}$	$\frac{22}{64}$

It is customary for us to ask the probability of getting a score of a certain size *or larger*, however, and these probabilities are given in the last column of Table 47. A score of 5 or larger includes scores of both 5 and 6; so the probability is $\frac{7}{64}$. A score of 4 or larger includes the three highest scores, and its probability of occurring is

$$\frac{1}{64} + \frac{6}{64} + \frac{15}{64} = \frac{22}{64}$$

We should not be amazed, therefore, if in a set of 6 trials as many as 4 correct judgments were given. This large a score can happen by guessing 22 times out of 64, or about 1 in 3. A score of 5 or larger can occur 7 times in 64, or once in 9 times. This is still not large enough to cause rejection of the null hypothesis. A score of 6 can happen by

chance once in 64 times. These are rather heavy odds against the hypothesis and lead us to place considerable credence in the student's assertion that he can discriminate the two tones. We should regard a score of 6 as significant, but we are not sufficiently sure to give up the null hypothesis entirely. We could conclude that 6 judgments, even when all are correct, are not enough for a genuine test of the matter.

We consider next a case with a larger number of trials; a set of 10 true-false test items to which a student gives one of two alternative responses, one right and one wrong. How many more than 5 items must he do correctly for us to reject the hypothesis that he knows nothing about the subject matter of the examination and that he is merely guessing at random? The probabilities corresponding to the three highest possible scores of 10, 9, and 8 are given in Table 48.

TABLE 48 — EXPECTED OCCURRENCES AND PROBABILITIES OF SPECIFIED NUMBERS OF CORRECT RESPONSES TO 10 TEST ITEMS

Number of correct responses	Times expected in 1,024 sets of responses	Probability of this number occurring at random	Probability of as many or more correct judgments
10	1	1/1,024	1/1,024
9	10	10/1,024	11/1,024
8	45	45/1,024	56/1,024

Following the same kind of reasoning as for the previous example, we should say that a score of 8 correct responses or higher could occur by random guessing 56 times in 1,024, or about 1 in 18 times. A score of 8 should thus not be very consoling to anyone. It might merely mean lucky guessing. A score of 9 or higher could happen only 11 times in 1,024, which looks much more favorable for something other than the no-knowledge hypothesis. There is only about 1 chance in 93 that so high a score could have occurred by guessing alone. And for the score of 10, which could happen only once in 1,024 times, we definitely reject the null hypothesis. We should feel free to reject it even when there are 9 correct responses out of 10, for then we are close to the limit arbitrarily chosen for a "very significant" deviation. This statement must not be generalized to cover cases where conditions are different from those specified; *i.e.*, a test of 10 two-response items and a student who, if he were completely ignorant, would respond at random. It is assumed that he would not be biased toward any

particular pattern or sequence of responses, as human beings frequently are!¹

Probability of Hypotheses Estimated from the Normal Curve.—In the previous illustrations, we actually counted up the total number of possible outcomes and also the number of times certain outcomes would be expected, and from these we obtained directly the probabilities that the null hypothesis was plausible. There are other instances, when the number of responses we deal with is quite limited, in which a similar counting of cases can be done and the probability of extreme deviations from chance can be derived. When the number of possible outcomes is not small, however, this counting of cases, or even algebraic computations of permutations and combinations, is much less efficient than other methods that will be described next.

In a certain elementary-psychology laboratory experiment, we have the problem to determine whether students can perceive from photographs whether or not a man has been convicted of crime. Pictures of 20 pairs of men matched for certain qualities are exhibited, and the student judges which of the two is the criminal. The null hypothesis calls for 10 correct responses, provided only that random guessing accounted for the score. How large an excess is indicative of actual perception or something other than chance?

To solve this problem, we do not resort to counting up the probabilities of as many as 20, 19, 18, etc., or more correct responses. Rather, we assume that each set of 20 judgments is a sample and that such samples would have a mean of 10, and a standard error of this mean will be the standard error of a frequency, which equals \sqrt{Npq} [see formula (30)]. We also assume a normal distribution of the samples of frequencies (see footnote 2 on page 162.) For this problem, N is 20, p is .5 and q is .5. The σ_f is therefore $\sqrt{20 \times .5 \times .5} = 2.236$. The distribution of these frequencies is shown in Fig 25, with a mean of 10 and a σ of 2.236. We are now ready to ask about the probability of a randomly determined score being as high as X or higher. For example, would a score of 14 be significantly in excess of the expected score of 10?

At first thought, this excess is 4 units above the mean of the distribution. But remember that a score of 14 is customarily one that occupies the interval from 13.5 to 14.5. A score of "14 or above" in this case therefore takes in all the normal curve above the point 13.5. It is a different matter to ask what is the area under the normal

¹ See Goodfellow, L. D., The human element in probability *J. gen. Psychol.*, 1940, **23**, 201-205.

curve above a point of 14.0 and to ask what is the area under the curve for a score of 14 or above. The deviation of the lower limit of this score from the mean is 3.5 units. Dividing this deviation by σ , which is 2.236, we have a z equal to 1.56. Going to the probability table (Table B) with this standard score, we find the area above the point 13.5 to be .0594. In other words, a score of 14 or above could occur about 6 times in 100 or about once in 17 times. A score of 14 is therefore not even significantly above chance expectation. A score of 15, which begins at 14.5, is 2.01σ above 10, and the probability of a chance score

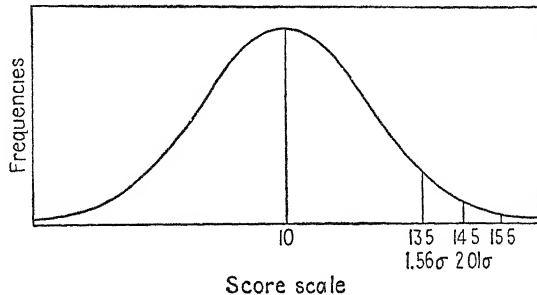


FIG. 25 —Standard-score distance from the hypothetical mean of 14, 15, and 16 correct judgments out of 20, when each judgment has an even chance of being right or wrong or complete ignorance

this high or higher is .0222, or is significant but not very significant, to use Fisher's standards.¹

A score of 16 is 2.46σ above the mean and so has only about 7 chances in 1,000 of being so large. When an individual's score is 16, therefore, we conclude that he was not merely guessing. And if all secondary cues, *i.e.*, cues not having to do with objective signs of criminality versus non-criminality in the photographs, were eliminated, we should conclude that this student can make this kind of discrimination. If, however, we had obtained 1,000 scores and only 7 (approximately) were this large, we should, on the basis of this much larger experience, revert to the null hypothesis. But from a single sample

¹ Fisher's standards or fiducial limits of 95 and 99 per cent, as defined in the preceding chapter, allow for 5 and 1 per cent (or proportions of .05 and .01) to be in the two tails at either end of the normal distribution. In the one case, the proportion in each tail is .025, and in the other case it is .005. In this chapter, we are concerned about deviations in one direction only, as a rule, and a significant deviation will have .025 or less beyond the point of significance, and a very significant deviation will have .005 or less in the tail. The levels "significant" and "very significant" are thus the same as before.

of 20 judgments, the statistical tests justify us in rejecting the hypothesis when the score is as high as 16.¹

How Large a Deviation Is Significant?—To return to the ESP problem, in 50 trials, when the probability of chance success is .20 and so the expected frequency is 10, the *SE* of the frequency is

$$\sqrt{50 \times .2 \times .8} = 2.82.$$

We could now test the plausibility of the null hypothesis in the face of different numbers of correct responses in excess of 10. But it might be more to the point to ask how large a score it would take to be significant and how large a score to be very significant.²

To be significantly in excess of 10, a score of X or larger could happen by chance only 2.5 per cent of the time (see the footnote on page 161). What point on the score scale comes at such a position? From the table, the z -score corresponding to this point is 1.96. This value times σ is 1.96×2.82 units on the score scale. This excess added to 10 gives us 15.3. Remembering that a score of 16 really begins at 15.5, we conclude that *at least* a score of 16 or higher is required to be significant of *anything* over guesswork. To be very significant, the tail probability is .005, the z -score is 2.576, and the excess is 7.26. This gives a point of 17.26 on the score scale. In terms of whole numbers, it requires a score of 18 or better to be very significant and to cause us to reject the null hypothesis. A score of 25 or better (above 24.5 on the scale) is 5.32σ above the mean, and there is about one chance in a million that so large an excess could occur by guessing alone. Such scores demand an explanation, but the explanation is not inevitably to be in terms of ESP unless other hypotheses have been adequately rejected.

How Large a Sample Is Necessary for Significant Deviations from Null Hypotheses?—We have already raised and answered the kind of question that asks for a given size of sample how large a discrepancy is

¹ Because N is as small as 20 here, however, we should recognize that Student's distribution of sampling applies rather than the normal distribution. Even with this revision, a score of 16 would be regarded as significant.

² This discussion assumes normal distribution of sample frequencies. When N is small and p deviates from .5 very far, this assumption no longer holds, for the distribution is skewed. Then one should determine the expected frequencies by applying the expansion of the binomial $(p + q)^N$. See Treloar, A. E., *Elements of statistical reasoning*. Ch. 12. New York: Wiley, 1939. Treloar suggests that the normal distribution is sufficiently approximated for practical purposes when the product Np is equal to or greater than 10 (or if q is less than p , then Nq is the criterion). According to this rule, our illustration barely qualifies as a case to which normal-curve reasoning applies.

necessary for significant and very significant deviation from a null hypothesis. Here we face a little different kind of question. We let our relative excess remain constant and ask how large N must be in order for that same size of discrepancy to reach the critical levels.

In a survey like the Gallup poll, for example, one would constantly be faced with the question of how large a sample to obtain; how many interviews to make; how many responses to a stimulus to record. That mere numbers in a sample as such are not sufficient to guarantee predictive ability was brought home to us decisively by the unhappy *Literary Digest* poll of 1936. Though the votes sampled ran into the millions, the voters who really determined the outcome of the presidential election were not adequately represented in the sample. A good poll sees to it that every kind of group of voters where group differences count at all are proportionately represented in the poll. When this is accomplished, it is surprising to the uninformed person how small a total sample can yield a valid predictive index. In other words, it is not so much enormous numbers that count as how the sample is made up.

Let us assume that our sample is properly made up with truly proportional representation. Let us assume an issue where majority vote is decisive. Our null hypothesis is then 50 per cent or a proportion equal to .50. We ask first how large a sample is needed to give us confidence that an obtained vote of 55 per cent in favor of the proposition means a majority sentiment in that direction and did not occur by random sampling from a population that is on the fence. If a discrepancy of as much as 5 per cent is to be significant in our accepted meaning of the word, 5 per cent must deviate as much as 1.96σ from the mean of a normal distribution. In terms of proportions, the deviation is .05, how large must σ_p be? Obviously it must be such that .05 is 1.96 times σ . σ_p is therefore equal to $.05/1.96$, which equals .0255. The formula we need is

$$N = \frac{pq}{\sigma_p^2} \quad (44)$$

We know p and q and σ_p already. Substituting them in the equation, we have

$$N = \frac{.5 \times .5}{.0255^2} = \frac{.25}{.0065025} = 384$$

to the nearest whole number. It is therefore a 19 to 1 bet that when a vote comes out with 55 per cent in favor of an issue in a sample of 384 that the population sampled is not evenly divided on the question.

But where much is at stake, we should not be satisfied with these odds against the null hypothesis. We might ask how many votes need to be sampled to assure us of a *very* significant deviation. In this case, the excess of .05 must be at 2.576σ from the mean. The σ_p must be $.05/2.576$, which equals .0194. Applying formula (44) to determine N , we have

$$N = \frac{.5 \times .5}{.0194^2} = \frac{.25}{.00037636} = 664$$

Thus, in a sample of 664 interviewees, a majority vote of 55 per cent would be regarded as very significant. The odds would be 99 to 1 that the sentiment of the population sampled is not evenly divided on the issue. And since the deviation is in the direction favoring the issue, we strongly expect future outcomes to be in the same direction, but we do not know by how much.

The sizes of samples just found are surprisingly small in view of the enormous populations that vote on national issues and whose sentiment they may be expected to estimate. The reason is that we have allowed a rather wide margin of .05 as the deviation from null hypothesis. In dealing with more vital issues, where close elections are concerned, excesses of .01 or less may be decisive. If we are interested in the sizes of sample required to give significant and very significant indications when the vote is .51 to .49, the SE of the proportion must be one-fifth as large as it was for a .55-to-.45 division. If σ_p is one-fifth as large, σ_p^2 is one-twenty-fifth as large. In this particular problem, the numbers to be substituted in formula (44) are now the same except that the denominator is one-twenty-fifth of its former size. This makes N twenty-five times as large as before.

For a deviation of .01 to be significant now, N must be 9,600, and to be very significant it must be 16,600, these numbers being 25 times 384 and 664, respectively. Samples of this size would give us great assurance, granting random sampling, that the sentiment is in the direction indicated. On many issues, of course, the sentiment is more unevenly balanced than .55 and .45. And, again, when we are interested in significance of changes in sentiment, we have a revision of our problem, for then we are dealing with differences among proportions, a kind of problem to which we now turn.

The Significance of Differences in Sampled Sentiment.—In a recent poll of radio-audience reactions,¹ out of 43 interviewees who

¹ Cantril, H., The rôle of the radio commentator. *Public Opinion Quarterly*, 1939, 3, 654-662.

were questioned, 72 per cent replied "Yes" to the question: "Do you find it easier to listen to news than to read it?" Although, in view of the one-sided expression of opinion, we might accept this as assuredly indicating majority belief that it is easier to listen to news than to read it, because of the small sample, we are challenged to make a test of statistical significance. On the basis of the null hypothesis, a 50-50 division of opinion, the *SE* of the proportion is

$$\sigma_p = \sqrt{\frac{.5 \times .5}{43}} = .0762$$

The excess of .22 over and above this hypothesis is 2.89 times the σ_p . Being in the "very significant" category, this result leads us to reject the hypothesis of evenly divided opinion. If a normal distribution is assumed, the probability of a deviation as large as .22 *in this direction* is given in Table B to be about .002, or 1 chance in 500, that so large a proportion as .72 could have occurred from sampling a 50-50 inclined population.

The author of the same survey goes on to say that when the data are fractionated into two parts according to higher and lower socio-economic status, the division of opinion is somewhat different. The higher group respond 10 "Yes" and 9 "No," whereas the division in the lower group is 20 and 4, respectively. We could test the null hypothesis for each group separately now, but it is quite obvious that for the higher group there is no significant deviation from a 50-50 hypothesis, and it looks quite favorable for a significant deviation for the lower group, although *N* is only 24. The proportions of votes are now .53 to .47 for the higher group and .83 to .17 for the lower.

We are more interested at this point in the question whether there is a significant difference between the two groups on the opinion about radio news. The higher group sample may actually represent a population with a more lopsided belief about the matter, whereas the lower-group sample may represent a population with a more even division of belief. We proceed with this problem as we did in the last chapter, where we determined the reliability of differences between proportions. For the sake of estimating the σ_{d_p} , we adopt as our hypotheses the observed proportions. The σ_p for the higher group turns out to be .115 and for the lower group, .076. The *SE* of the difference, if no correlation between the two groups is assumed, is .138. The ratio of the difference (.83 - .53) to the σ_{d_p} is .30/.138, which equals 2.17. As a *t* ratio, this is just barely significant.

Let us look at the result in another manner with the aid of Fig. 26. This normal curve represents the distribution of differences among proportions of "Yes" responses for many pairs of samples from the two populations (higher and lower socioeconomic groups). The mean is placed at a difference of zero. A difference as large as the obtained one, .30, then occurs at the point indicated, 2.17σ above the mean, or on the side of the positive differences. On the hypothesis of zero difference in the populations, a deviation as large as $+.30$ or larger could occur 15 times in 1,000, or about once in 67 times.

To put the interpretation in somewhat new terms, we may say that the proportion of .015 under the tail of the normal curve *above* the point of difference $+.30$ gives us the probability that the true

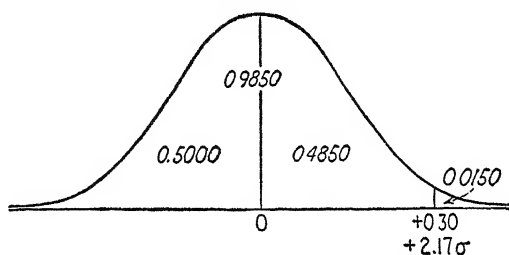


FIG. 26—Distribution of hypothetical samples of differences in proportions of higher and lower socioeconomic groups who report radio news preferable. The obtained difference is 2.17σ above the hypothetical mean of zero.

difference is a negative one, and the remainder of the area *below* the point, or .985, gives us the probability that the true difference is positive. The odds are therefore .985 to .015 that the true difference is positive, or about 67 to 1. What practical measures one might wish to adopt as a consequence of this finding would depend upon his own inclination to accept odds of this nature and the urgency of the outcome of planning one course of action or another.¹

As another illustration, let us consider some data on the marital status of feeble-minded (low *IQ*) men as compared with men of normal *IQ*. Table 49 furnishes us with some data, for 206 men whose *IQ*'s were in the 60's were matched for age with 206 men whose *IQ*'s were near 100.² From these data, we find that at certain ages in the twenties, when they were compared, .539 of the average men and .408 of

¹ With such small samples as we have here, normality of distribution of differences is, of course, questionable. We have carried the illustration through, nevertheless, to show how larger samples may be treated.

² Baller, W. R., A study of the present status of adults who were mentally deficient *Genet Psychol Monogr*, 1936, 18, 165-244.

the mentally defective were married. Is this difference significant? The SE of the difference between proportions is estimated to be .049. The difference itself is 2.67 times the σ_{d_p} . Such a t ratio is regarded as very significant, for a difference as large as .131, which was obtained, could have occurred by sampling error in *either* direction (plus or minus .131) less than once in a hundred times.

TABLE 49.—A COMPARISON OF MEN OF NORMAL IQ WITH FEBBLEMINDED MEN WITH RESPECT TO MARITAL STATUS

Marital status	Normal	Feeble-minded	Both
Married	111	84	195
Unmarried	95	122	217
Total . . .	206	206	412

To apply the new interpretation suggested above, we may say that the area under the tail of the normal curve above $+.13$, which is .004, gives the probability of a true difference in the opposite direction, and .996 is the probability of a true difference in the obtained direction.

CHI SQUARE

Chi Square as a Test of Deviation from Null Hypothesis.—Some statisticians are more inclined to employ the chi-square technique to this kind of problem. In this procedure, we also make a null hypothesis and determine how likely it is that our sample could have diverged from this hypothesis as much as it did, had chance factors alone been operating. In the data on marital status of two different groups of men, we should assume that they really come from the same population in this respect or that the two populations that they represent are alike with regard to marital status.

In Table 49, we find that if we take the two groups combined, 195 were married and 217 were not. The proportions are .4733 and .5267, married and unmarried, respectively. If the two populations are actually alike, both should have the same ratio of married to unmarried, or .4733 to .5267. The expected numbers of married and unmarried in a sample of 206 would be 97.5 and 108.5. The null hypothesis therefore calls for these frequencies, which we term *expected frequencies* and symbolize by f_e . The obtained frequencies are symbolized by f_o . In Table 50, the expected frequencies are listed for the two groups. With the same totals in the two groups, 206, the expected frequencies will be the same. This will not be the case in all problems of this sort.

Having the expected frequencies f_e , we now ask whether the observed frequencies f_o deviate from them sufficiently to cause us to reject the hypothesis of no difference. For each of the four cells of the table, we determine the discrepancy $f_o - f_e$. These are listed in Table 51. It

TABLE 50 —THE EXPECTED NUMBERS OF MARRIED AND UNMARRIED MEN IN THE NORMAL AND FEEBLEMINDED GROUPS HAD THERE BEEN NO DIFFERENCE BETWEEN THE TWO

Marital status	Normal	Feebleminded	Both
Marned	97.5	97.5	195
Unmarried	108.5	108.5	217
Total	206	206	412

TABLE 51 —DISCREPANCIES BETWEEN OBTAINED AND EXPECTED FREQUENCIES IN TABLES 49 AND 50

Marital status	Normal	Feebleminded
Marned	13.5	-13.5
Unmarried	-13.5	13.5

will be seen that, except for algebraic sign, they are all numerically the same. This will be true of all fourfold tables of frequencies of this sort, whether the two groups compared have the same total numbers of cases or not. This fact can be used to give us short cuts in computation, as we shall see later.

The solution of chi square calls upon us to square each discrepancy, divide this by the corresponding f_e , and sum all the ratios. In terms of a formula, it is

$$\chi^2 = \sum \left[\frac{(f_o - f_e)^2}{f_e} \right] \quad (45)$$

where the symbols are as explained above.

The square of the discrepancy, 13.5, is 182.25. In two cells, this is to be divided by 97.5, which yields 1.87. In the other two cells, it is to be divided by 108.5, which yields 1.68. Summing twice 1.87 and twice 1.68, we get 7.1 as the value for χ^2 . This number now stands for the total amount of discrepancy between hypothesis and observation. Chi square can be small enough to allow us to accept the null hypothesis or to retain it with some doubt, or it can be large enough to lead us to reject the hypothesis with moderate or with positive assurance. Like student's ratio, it can be interpreted as being significantly

or very significantly large, *i.e.*, of being so large that chance factors alone could be responsible only once in 20 times, or once in 100 times, as the case may be.

Degrees of Freedom.—Tables of chi square (see Table E) enable us to decide the matter. But we must know the number of degrees of freedom n before we can use the table. In a fourfold table such as we have here, there is only 1 degree of freedom.

Is our chi square of 7.1 significant? Table E shows that when $n = 1$, the largest chi square given is 6.635. Right above this is the probability of .01, which means that a chi square as large as 6.635 could occur by chance alone only once in 100 times. Our chi square of 7.1 is larger than 6.635 and therefore could occur in the same manner less than once in 100 times. We therefore regard it as very significant and reject the hypothesis of no difference between the two groups.

In a fourfold-table problem such as this, since the discrepancy is the same for all cells, the formula for chi square can be written

$$\chi^2 = (f_o - f_e)^2 \sum \left(\frac{1}{f_e} \right) \quad (46)$$

That is, chi square equals the square of the common discrepancy times the sum of the reciprocals of the four f_e 's. As applied to the marital-status problem

$$\begin{aligned} \chi^2 &= 13.5^2 \left(\frac{1}{97.5} + \frac{1}{97.5} + \frac{1}{108.5} + \frac{1}{108.5} \right) \\ &= 182.25 (.01026 + .01026 + .00922 + .00922) \\ &= 182.25 \times .03896 \\ &= 7.10 \end{aligned}$$

Chi Square When Frequencies Are Small.—When any of the cell frequencies are small (less than 50), it is good practice to make allowances that are theoretically necessary. The allowance made is to deduct .5 from each one of the discrepancy values. This is known as Yates's *correction for continuity*.¹ It is needed for the same reason that earlier in the chapter we treated the deviation of a certain score from hypothesis not as a point on the scale but as the lower limit of an interval, a half unit below the score value (see page 160). In brief, the theory of sampling that uses a hypothetical distribution curve assumes a smooth, continuous variation, whereas scores and frequencies jump by discrete units or intervals. This point is not of serious consequence when the sample is large but becomes relatively so when the sample is small.

¹ Snedecor, G. W., *Statistical methods*. Ames, Iowa: Collegiate, 1937. P. 161

As an example of a small sample, let us use again the data on responses to radio newscasts.¹ First, let it be said that we can apply the chi-square test to only two frequencies as well as to four or more. Of 43 people interviewed, 72 per cent, or 31, replied "Yes" and 12 replied "No." On a hypothesis of 50-50 division of opinion, we should have expected 21.5 "Yes" responses and 21.5 "No" responses. The discrepancies $f_o - f_e$ are both numerically 9.5. But when the correction for continuity is applied, these values shrink to 9 even. The discrepancy squared is 81. Divided by f_e in each case (now the same, 21.5) the ratio is 3.767. The sum of two of them is 7.534, which is chi square. With 1 degree of freedom, this is found to be a very significant deviation from hypothesis, which agrees with our earlier conclusion about the same data.

Here we have made a more strict test of significance, because the correction for continuity was applied. We could have applied the same correction in the earlier test, in that the excess that was given before ($9.5/43$ was stated to be .22) should have been replaced by one of .21 (*i.e.*, $9/43$). The general conclusion would have been the same, as it is for the chi-square test, but when the criterion of significance, t ratio or chi square, is near the border line, the correction for continuity may make an important difference, particularly when the sample is very small.

With a twofold table, as in the last illustration, the formula for χ^2 can be reduced to

$$\chi^2 = \frac{2(f_o - f_e)^2}{f_e} \quad (47)$$

We actually followed the steps implied by this equation in the solution of χ^2 in the last paragraphs.

Chi Square Computed from Proportions.—If the data are given in terms of proportions of cases rather than in terms of frequencies in a fourfold table, a chi square can be computed directly from the proportions. In Table 52, the data on marital status are given throughout as

TABLE 52.—THE DATA ON MARITAL STATUS REDUCED TO PROPORTIONS AND A GENERALIZED FOURFOLD TABLE EXPRESSED IN SYMBOLS

	Normal	Feeble-minded	Both		First group	Second group	Both
Married	269	.204	473	First quality	α	β	p
Unmarried . .	.231	.296	.527	Second quality.	γ	δ	q
Both . .	500	.500	1 000	Both . .	p'	q'	1 000

¹ Cantril, *op. cit.*, p. 658.

proportions of the total number of cases (412). Corresponding to the table of proportions is presented another in which letter symbols are substituted for the proportions. The formula for chi square under these circumstances is

$$\chi^2 = \frac{N(\alpha\delta - \beta\gamma)^2}{pq p' q'} \quad (48)$$

where the symbols are as explained in Table 52. For this particular problem

$$\begin{aligned} \chi^2 &= \frac{412[(.269)(.296) - (.204)(.231)]^2}{(.5)(.5)(.473)(.527)} \\ &= \frac{(412)(.0325)^2}{.06231775} \\ &= \frac{.435175}{.06231775} \\ &= 6.98 \end{aligned}$$

which checks fairly closely with the χ^2 computed before from the frequencies.

In the special case where the two experimental groups contain the same number of cases—in other words, when $p' = q' = .5$ —the formula reduces to

$$\chi^2 = \frac{N(\alpha - \beta)^2}{pq} \quad (49)$$

In this problem, then

$$\chi^2 = \frac{(412)(.065)^2}{(.473)(.527)} = \frac{1.7407}{.249271} = 6.98$$

Chi Square in Larger Tables of Frequencies.—The use of the chi-square test is not by any means confined to a comparison of two or four frequencies. It can be employed to tell whether any set of observed frequencies deviates significantly from almost any hypothetical set of frequencies, except when theoretical frequencies become very small. To illustrate one more application of chi square, this time to a table of six frequencies, let us consider some more survey-of-opinion data.¹ This time the question was whether the radio listener agreed with the opinions expressed by a certain radio commentator, and the responses were tabulated as "Agree," "Disagree," or "Doubtful." The survey was made in two cities and we have the numbers responding in each way in both of them. The results are listed in Table 53.

¹ Cantril, *op. cit.*

TABLE 53—A CHI-SQUARE SOLUTION IN A TWO-BY-THREE TABLE OF DATA ON OPINIONS EXPRESSING AGREEMENT OR DISAGREEMENT WITH A CERTAIN RADIO COMMENTATOR

Categories of response		Opinions in Syracuse	Opinions in Columbus	Both	Per cent
Agree		73	22	95	54 0
Disagree		9	4	13	7 4
Doubtful		41	27	68	38 6
Totals		123	53	176	100 0

f_e Expected frequencies		$f_o - f_e$ Discrepancies		$(f_o - f_e)^2$ Discrepancies squared		$\frac{(f_o - f_e)^2}{f_e}$ Ratios	
Syracuse	Columbus	Syracuse	Columbus	Syracuse	Columbus	Syracuse	Columbus
66 4	28 6	+6 6	-6 6	43 56	43 56	0 66	1 52
9 1	3 9	-0 1	+0 1	0 01	0 01	0 00	0 00
47 5	20 5	-6 5	+6 5	42 25	42 25	0 89	2 06
123 0	53 0	0 0	0 0	.	.	1 55	3 58

In order to determine the expected frequencies f_e for the three categories of response in the two cities, we first determine the percentage in the two cities combined that respond in each way. These are given as 54.0, 7.4, and 38.6 per cent for the three responses "Agree," "Disagree," and "Doubtful," respectively. The f_e 's for the Syracuse group were found by taking these percentages of 123. The f_e 's of the Columbus group were found by taking these percentages of 53.

From here on, the work is just as before and is systematically arranged in Table 53. Checkings of percentages, f_e 's, and of the discrepancies $(f_o - f_e)$, are made by summing their columns. The sums of the discrepancies should equal approximately zero. The chi square in this case is 5.13. Its degree of significance is yet to be determined. The number of degrees of freedom is 2 in this problem. In general, the number of degrees of freedom is the number of rows minus 1 (2 in this problem) times the number of columns minus 1 (1 in this problem). For 2 degrees of freedom, Table E tells us that it requires a χ^2 of 5.991 to be significant. Our χ^2 of 5.13 lacks significance; so we say that there is not sufficient reason for doubting that the two populations sampled are alike on the question at issue, though there are

less than 10 chances in 100 that a chi square this large could have arisen by chance

But the student may recall what was said previously about the need for correction for continuity. That need surely exists in this problem, in which four of the six f_e 's are less than 50. Had we made this correction, χ^2 would have been even smaller. Since it was not significantly large without the correction, it certainly would not be with the correction; so we need not recalculate χ^2 with the correction

As a matter of fact, it is a good rule not to compute a χ^2 at all for a table of frequencies in which any f_e is less than 5. We have ignored the rule here for the sake of an illustration, for we have one such f_e in this problem. The way out is to condense the table to a smaller number of columns or rows, or both, eliminating the smallest frequencies. If we had combined the "Disagrees" with the "Doubtfuls," we should have observed frequencies of 50 for Syracuse and 31 for Columbus and expected frequencies of 56.6 for Syracuse and 24.4 for Columbus in the new combined cells. The size of chi square is reduced in this process; but so is the number of degrees of freedom, so that a smaller χ^2 is required for the same level of significance.

Chi Square in Testing the Hypothesis of Normal Distribution.—One convenient use of chi square is in testing whether or not a set of observed frequencies in a frequency distribution could probably have arisen from a normally distributed population. The fact will be barely mentioned here, however, for the reason that the usefulness of this application of χ^2 is restricted to the rare problem in which it is required. The procedure is carried out in much the same manner as with frequencies in this chapter. Expected frequencies are estimated as was illustrated in Ch. VI, particularly in Table 25 (page 81). The discrepancies between observed and expected frequencies are squared, divided by the f_e 's, and these ratios are summed to give χ^2 . The number of degrees of freedom is the number of class intervals less three. At the tails, where f_e 's are small (less than 5), two or more class intervals should be grouped together as was suggested in recent paragraphs for other data. The interpretation of the result is made according to precedents already repeated, and the hypothesis of normality is accepted or rejected according as χ^2 is small or large.

Exercises

1. Suppose that we ask an observer to arrange a series of weights in rank order from lightest to heaviest, the differences being very small. If he places them in perfect rank order, what is the probability that he could have done so by sheer guessing? No matter how many weights ranked, there is only one correct way of doing

this. The total number of ways the observer could have arranged each number of weights is given below

Number of weights	3	4	5	6	7
Number of orders	6	24	120	720	5,040

Which perfect orders would be regarded as "not significant," "significant," and "very significant"? State the probabilities of perfect orders by chance

2 An observer knows that he will hear one of three similar speech sounds. He is given the three in chance order in a total of 30 trials. How many correct judgments must he give before we regard his success as significant and as very significant?

3 Suppose that the observer in Exercise 2 were given 48 trials. How large a score is significant, and how large a score is very significant?

4 A certain examination includes 40 items, each item with four alternative responses. How large a score must a student earn before you feel that he probably knows something about the content of the examination? Before you feel that he undoubtedly knows something about it? Would you feel absolutely sure that he knows something about the content if he made a score of 35? Discuss.

5 In a test of five-response items, how many items would you need to include in order to feel sure that a score of 30 per cent right indicates knowledge of the content? How large must the test be if a score of 25 per cent right is to indicate knowledge beyond a reasonable doubt? Tell how you have interpreted "sure" and "beyond reasonable doubt."

DATA V.—NUMBER OF PERSONS IN TWO GROUPS, DEPRESSED AND NOT DEPRESSED IN TEMPERAMENT, WHO RESPONDED IN EACH OF THREE CATEGORIES TO THE QUESTION, "WOULD YOU RATE YOURSELF AS AN IMPULSIVE INDIVIDUAL?"

Group	Yes	?	No	Totals
Depressed	72	45	133	250
Not depressed	106	35	109	250
Totals	178	80	242	500

6 Is there a significant difference in Data V between the numbers of "Yes" responses? Present statistical proof

7 Is there a significant difference between the two groups in the number of "?" responses? Explain

8. Is there a significant difference in the two groups with regard to all three response categories taken together? Determine this by computing chi square.

9. State a number of null hypotheses that might be applied to Data W.

DATA W.—NUMBERS OF TWO GROUPS DIFFERING IN ABILITY WHO PASSED A CERTAIN TEST ITEM

Group	High group	Low group	Both
Passed	62	48	110
Failed	38	52	90
Both	100	100	200

10 Do both groups together in Data W show a significant deviation from a chance situation of passing and failing? Explain.

11 Is there a significant difference between the high and low group in terms of the numbers passing the item? Explain Can you predict from this result whether there would be a significant difference between numbers of failures in the two groups? Explain.

12 Find a chi square for Data W in as many ways as you know how Interpret your results

13. In Data V , combine the "Yes" and "?" responses, and compute chi square for the fourfold table. Compare your results with those in Problem 8

CHAPTER X

PREDICTION AND ERRORS OF PREDICTION

One of the most important fruits of scientific investigation and one of the most exacting tests of any hypothesis is the ability to make predictions. So important is this topic that it deserves at least a chapter devoted to it. Particularly is this true for the reason that statistical reasoning is basic to all predictions. Statistical ideas not only guide us in framing statements of a predictive nature but also enable us to say something definite concerning how trustworthy our predictions are—about how much error one should expect in the phenomenon predicted. The practical significance of this cannot be questioned. The significance even for the scientific investigator is too often unrecognized and forgotten.

One can find amateur prognosticators for almost any kind of event on every hand. Little note is made of the success or failure of their predictions. A few successes are sufficient basis for vindication of the prophet, and many failures are quickly forgiven and forgotten. The old adage "Where ignorance is bliss 'tis folly to be wise" must have been invented to fit this particular situation. On the other hand, the psychologist or educator who falls short of perfect predictions is often immediately condemned and his further predictions thought to be discredited. The average uninformed person is somehow partial to vague and "magical" means of prediction, and he can readily overlook their shortcomings, whereas he will not tolerate the statistically hedged prediction that also yields to him a more exact knowledge of its limitations. If he could only realize how poor the predictions of the amateur prophet actually are, he would perhaps have a more ready respect for the scientific prediction of events in human affairs. It is the purpose of this chapter to illustrate the kinds of predictions the statistically oriented investigator makes and how he not only does not blind his eyes to his failures but brings them clearly into the light.

General Types of Prediction.—Although in this volume we have generally emphasized measurement, we have had to recognize from time to time that complete measurements cannot be made and that data are sometimes obtained as merely classified in categories. The

latter type of data we recognize as *enumeration data* rather than as measurements. It is a matter of assigning attributes to cases rather than quantitative evaluations on a linear scale, for example, identifying individuals as to sex, race, political party, or criminality. Although such data are not allocated to linear-scale positions, we can still make predictions from them and of them from other information. We thus have four cases of predicting:

1. Attributes from other attributes—as when we predict incidence of criminality from sex, race, or religious creed.

2. Attributes from quantitative measurements—as when we predict criminality from scores on tests of ability or of behavior traits.

3. Measurements from attributes—as when we predict probable test scores from sex, socioeconomic status, or marital status.

4. Measurements from other measurements—as when we predict achievement in school from *IQ*-test scores.

General Ways of Evaluating Accuracy of Prediction.—Predictions are obviously sound if they prove to be correct. The degree of correctness depends upon *how often* or *how nearly* we hit the mark. In the case of predicting attributes, our success can be numerically indicated in terms of the percentages of “hits.” But a more accepted way among statisticians is to ask how much better our predictions are than if we had not used the information we have—in other words, if we had not tried to predict one thing from the knowledge of another but merely from a knowledge of the predicted population itself. A more crude way of saying it would be to ask how much better our predictions are than guesswork. But this does not mean *pure* guesswork, as we shall see later.

In predicting measurements, whether from attributes or from other measurements, we ask a similar question. But whereas in predicting attributes for cases, we work in terms of the *number* of hits or misses, since we are dealing with enumeration data, in predicting measurements, we work in terms of *how far* on the average we have missed the mark. We compare this average deviation between fact and prediction with the average of the errors we should make without using the knowledge we did as a basis of prediction.

Let us see in a preliminary way what this means. We can predict that a student's mark in a course will be somewhere in the range from A to F inclusive, and most probably it will be a mark of C, which more students earn than any other mark. This prediction is made without knowledge of the student's scholastic-aptitude score, and its margin of error is measurable in terms of the standard deviation of the distribu-

tion of marks of all students. If we used knowledge of the students provided by aptitude-test scores, we should predict some to earn marks higher than C and some lower than C. The average of our deviations between prediction and fact will now be smaller than the standard deviation of the distribution of all marks. The difference between these averages of deviations tells us how much the knowledge of aptitude scores has improved our predictions.

PREDICTING ATTRIBUTES FROM OTHER ATTRIBUTES

Predictions Can Be Made in Both Directions.—As our first example of prediction of attributes from other attributes, let us consider the data in Table 54. Here we have the numbers of persons in a "depressed" group who responded by saying "Yes," "?" and "No" to the question, "Would you rate yourself as an impulsive individual?" and also the number of a group described as "not depressed." The individuals in these two categories are the highest and lowest quarters of a sample of 1,000 students who were ranked in terms of a provisional scoring on a personality inventory. Table 54 provides us with two prediction

TABLE 54.—DISTRIBUTION OF RESPONSES TO THE QUESTION "WOULD YOU RATE YOURSELF AS AN IMPULSIVE INDIVIDUAL?" AS GIVEN BY TWO EXTREME GROUPS OF STUDENTS

Group	Response			
	Yes	?	No	Total
Depressed	72	45	133	250
Not depressed	106	35	109	250
Both	178	80	242	500

problems. We can attempt to predict the verbal response to the question, knowing whether the person is in the depressed or not depressed group; or we can attempt to predict the group to which a person belongs, knowing what response he has made. Let us take the prediction of verbal response first.

The Principle of Maximum Likelihood.—Considering first the depressed group by itself, we find that the largest number of them respond with "No." Taking each member of the depressed group as he came along, we should predict for him the response "No." If all 250 came up for inspection, we should be correct 133 times out of 250, or 53.2 per cent of the time. For smaller samples from the same depressed population, we should expect a similar ratio of correct predictions.

This illustration sets the pattern for all predictions of attributes from attributes. The prediction always observes the *mode* or most frequent attribute in the segment of the population chosen at the moment. For the not depressed group, the mode is also at the response "No"; hence that is our prediction also for them, and our percentage of accuracy is 43.6 per cent, not so high as before but higher than if we had predicted either "Yes" or "?" for this group. Such predictions follow the *principle of maximum likelihood or maximum probability*. Either a depressed or a not depressed person in this population is more likely to respond "No" than anything else; so that is our prediction.

The Forecasting Efficiency in Predicting Attributes.—How good are these predictions? Since we have predicted the same response for both depressed and not depressed individuals, we suspect that knowing to which group the person belongs helps us little if any to predict his response. A comparison of the percentages of correct predictions, however, tells us that we can be more sure of our prediction of "No" if the person is depressed than if he is not. But no matter from what group the person comes, our prediction is the same; so it is as if we could make no use of the knowledge of his group affiliation for this purpose.

Let us compare the number of successes of prediction made with and without knowledge of group affiliation. Taking both groups combined, we should predict for each person at random the response "No," and we should be correct 242 times in 500, or 48.4 per cent. In the two groups predicted separately, we found successes of 133 and 109, which combined give us 242 correct hits, or 48.4 per cent. We have thus gained no more accuracy in predicting responses from a knowledge of group affiliation than we could attain without this knowledge. The *forecasting efficiency* in predicting response from knowledge of group is therefore just zero. The work of calculating forecasting efficiency may be seen more clearly if summarized as in Table 55.

TABLE 55.—PREDICTIONS OF RESPONSE FROM KNOWLEDGE OF THE GROUP MEMBERSHIP

Group membership	Predicted response	Number correct	Per cent correct
Depressed	No	133	53 2
Not depressed.	No	109	43 6
Total.		242	48.4
Correct without knowledge		242	48 4
Excess with knowledge.		0	0.0

The second prediction problem here is to reverse matters and predict group membership from knowledge of the response. All persons responding "Yes" we should predict to be members of the not depressed group, since 106 actually are, as compared with 72 who are not. Again the *modal* attribute is our prediction. For those responding "?" the prediction is membership in the depressed group, and so also for those responding "No." The percentages of correct predictions are given in Table 56 for each response and for all combined. Altogether, there are 284 correct predictions, or 56.8 per cent. Without knowledge of which response each person made to the question, but with knowledge that half the total population are depressed and half are not, our expected number of chance successes is 250. Our predictions *with* knowledge of responses yielded an excess of 34 or a *forecasting efficiency* of 13.6 per cent. We can say that our predictions with knowledge of response to the question is 13.6 per cent better than those made without this knowledge would be.

TABLE 56—PREDICTIONS OF GROUP MEMBERSHIP FROM KNOWLEDGE OF VERBAL RESPONSE TO THE QUESTION

Response	Predicted group	Number correct	Per cent correct
Yes	Not depressed	106	59.6
?	Depressed	45	56.3
No	Depressed	133	55.0
Total		284	56.8
Correct without knowledge		250	50.0
Excess with knowledge		34	13.6

Prediction Not Equally Good in the Two Directions.—It is now well apparent that we can predict successfully group membership from knowledge of responses in this problem, whereas we cannot predict response from knowledge of group membership. It is not always true, as it is here, that successful prediction is possible in one direction and *entirely* impossible in the other, but it is a quite common finding that prediction is better in one direction than in the other when two variables are concerned. It will often clarify thinking about predictive problems to keep this fact in mind. It is sometimes assumed by the uninformed that if *A* can be predicted from *B*, *B* can, in turn, be predicted from *A*. Such an assumption is likely to lead the unwary investigator into logical and practical difficulties when it is seriously wanting in applicability. This is a more serious matter in dealing with attributes than in

dealing with measurements, for in the latter case the predictability of one measured trait *A* from a measured trait *B* cannot be very divergent from the predictability of *B* from *A*.

PREDICTING AN ATTRIBUTE FROM MEASUREMENTS

We sometimes wish to decide on the basis of measurements that we know, whether an individual should be expected in one category as having a certain attribute or whether he should be expected in another. Sometimes it is a matter of making placements in different categories in order that the individual may expect a better consequent adjustment or greater satisfaction. Such is the case when we attempt to predict success or failure for persons for whom we know certain test scores. This problem was recently solved in principle by Guttman.¹ Here we

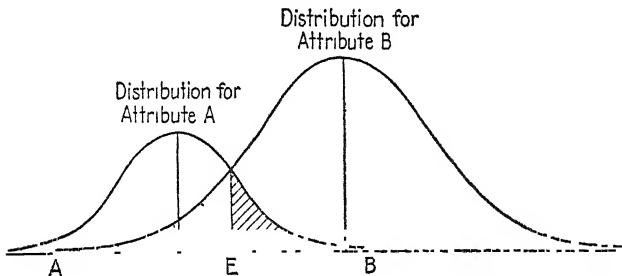


FIG. 27—Distributions of two 'better' groups possessing two distinguished attributes, *A* and *B*, when measured on the same scale. The aim is to predict for each person his attribute from knowledge of his score. For those with scores above point *E* we predict attribute *B*, for those below, attribute *A*.

shall attempt to provide some workable procedures whereby such predictions can be made and their relative accuracy determined.

Critical Points Dividing Distributions.—In Fig. 27, we have two populations, each normally distributed but differing in mean, standard deviation, and in *N*. For the main purpose of discriminating between the two populations, normality need not be assumed, but certain solutions are made easier if we do so. We wish to find a score on the scale of measurement that will give us the maximum accuracy of prediction, so that we may say of an individual whose score is higher than that point that he is probably a member of the upper group and of an individual whose score is lower than that point that he is probably in the lower group and in so predicting, make the minimum number of mistakes. Let us call that critical point *E*.

¹ The prediction of personal adjustment. New York: Social Science Research Council, 1941. Pp. 271*f*.

According to Guttman's solution, point *E* comes on the scale where the two distributions have equal ordinates—in other words, where the two curves intersect (see Fig. 27). At this point, persons with scores of this value are equally likely to be members of either group. Above this point, at any score there is greater likelihood that the person belongs in the upper group than that he belongs in the lower group. Below this point, at any score, there is a greater likelihood that the person belongs in the lower group. The terms *upper* and *lower* here apply only to relative position on the measuring scale. The two distributions are divided according to two qualities or attributes, and it is possession of those attributes that we are trying to predict. As we proceed along above point *E*, the probability that we are correct in our prediction increases, since the ratio of the individuals having attribute *B* to the number having attribute *A* keeps increasing. At point *B*, which is the upper limit of the range of the *A* group, and above *B* we should have absolute certainty of prediction so far as these particular populations are concerned. Likewise, below point *A*, where the upper distribution ends, we should be absolutely certain that no case possesses attribute *B*. But if the two populations are taken as wholes, the shaded portions stand for the proportions of individuals incorrectly predicted. The cross-hatched section represents the *A*'s wrongly predicted to be *B*'s, and the stippled section represents the *B*'s wrongly predicted to be *A*'s. All the *B*'s above point *E* are correctly predicted, and all the *A*'s below point *E* are correctly predicted. It is on the basis of these numbers of correctly and incorrectly predicted cases that we can judge the forecasting efficiency, as we shall see later. First, let us see how point *E* can be determined.

How to Locate a Critical Division Point between Distributions.—

There are several procedures by which point *E* can be estimated. The most accurate one, which assumes normal distributions, requires the solution of a complicated quadratic equation and is prohibitive in the amount of labor involved. When distributions are lacking in normality, too, such a procedure cannot well be employed. The writer will describe two methods that will yield a result sufficiently accurate for most practical purposes, one a graphic method and the other an arithmetical solution.

As illustrative material, let us use the data in Table 57. A large group of students were given the same comprehensive final examination in freshman English. Each instructor was at liberty to use the scores in this examination along with other measurements as he saw fit in deriving a final mark in the course for his students. Taking all the

marks collectively, for all students receiving a mark of F, a frequency distribution of their examination scores was set up. The same was done for students receiving marks of D, C, B, and A. These are the five distributions listed in Table 57 and shown graphically in Fig 28. The amount of overlapping in ability as represented by examination scores among these five groups is noteworthy, but it probably represents a not unusual situation where marks are determined in the customary

TABLE 57—DISTRIBUTIONS OF SCORES IN A GENERAL ENGLISH EXAMINATION MADE BY STUDENTS RECEIVING VARIOUS MARKS IN THE COURSE

Scores	A	B	C	D	F
180-189	1				
170-179	1	1			
160-169	5	7	1		
150-159	7	13	3		
140-149	2	26	10	1	
130-139	2	34	24	5	1
120-129	0	40	39	7	0
110-119	1	21	81	13	3
100-109		19	89	28	4
90- 99		4	81	29	9
80- 89		1	42	46	8
70- 79			16	29	11
60- 69			5	20	9
50- 59				6	11
40- 49				1	5
30- 39					3
20- 29					0
10- 19					0
0- 9					1
Sums	19	166	391	185	65

manner. However that may be, let us say that students receiving F's are, in the judgment of the teachers, failing students, and those receiving D's are D students, etc. These five categories represent five attributes as judged by these instructors. Let us take as our rather artificial problem the task of predicting what attribute will be assigned to students making certain scores in the examination. The same principles apply in other instances where it would be more important to predict attributes from scores.

A Graphic Method of Locating the Critical Point.—When the overlapping distributions are plotted as in Fig 28, if they are fairly regular in contour, one can immediately locate the points at which two distributions intersect. Distributions for attributes F and D intersect just below a score of 60; more exactly, by inspection, at 57 or 58. In this approach, it would be well to locate the point between two whole numbers, because scores are obtained in whole numbers. In this case, we should predict an F for students making a score of 57 or lower and a mark of D for those making a score of 58 or above (at least up to the critical point between D and C). Between D and C, the critical

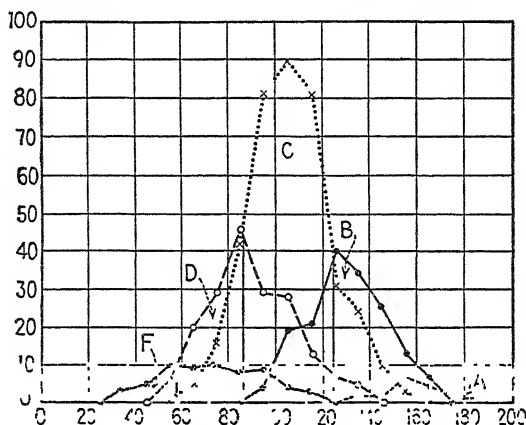


FIG. 28.—Distributions of students receiving marks of A to F in freshman English, in scores received in a common final examination.

point, by inspection, seems to be at about 87, probably on the lower side. Thus, for scores 58 through 86, we should predict a mark of D. The next critical point seems to come between 124 and 125. The prediction of a C arises for scores 87 through 124. The critical point between B and A is almost impossible to determine but seems to lie in the region of 170 to 175. The small number of A's make any solution of this kind uncertain.

Should any overlapping distributions be irregular in contour, particularly in the neighborhood of the intersection point, then, if the data are not too limited and if the smoothing required is rather obvious, it would be well to resort to smoothing before the point of intersection is sought (see page 23 for a description of smoothing procedures). Furthermore, if the two distributions are fairly normal in contour and if the results seem important enough to justify the labor involved, one might determine the best-fitting normal frequencies (see

page 81 for directions) before plotting the curves from which the point of intersection is determined.

An Arithmetical Approximation of the Critical Point.—It has been customary in setting up frequency distributions to assign the frequency for any class interval to the midpoint score value of that interval. The intersection of two distribution curves at the point where they have equal frequencies will practically always occur between two midpoints of neighboring intervals. By inspecting the frequencies of two overlapping distributions, we can readily narrow down the point of crossing to some place between two such midpoints. For example, in Fig. 28, the crossing of the curves for groups F and D is obviously between the midpoints 54.5 and 64.5, at which points the frequencies are plotted. This part of the two distributions is enlarged in Fig. 29. If we connect by vertical dotted lines the two pairs of frequencies at these midpoint values, we produce two similar triangles. By geometric reasoning, the altitudes of these triangles are directly proportional to their bases. If we let the dotted lines be the bases of these triangles, their altitudes are precisely the distances of point *E* from the two midpoints. Their bases are known quantities, for they are the differences between frequencies at those midpoints. In terms of an equation, the proportionality can be expressed as follows:

$$\frac{X_2 - E}{E - X_1} = \frac{f_{22} - f_{12}}{f_{11} - f_{21}} \quad (50)$$

where X_2 = higher of the two midpoints.

X_1 = lower of the two midpoints.

f_{22} = frequency of the upper group at midpoint X_2 (the higher midpoint).

f_{12} = frequency of the lower group at midpoint X_2 .

f_{11} = frequency of the lower group at midpoint X_1 .

f_{21} = frequency of the upper group at midpoint X_1 .

All the symbols in equation (50) are labeled in Fig. 29.

To apply the formula to the solution of the critical point between the D's and F's

$$\frac{64.5 - E}{E - 54.5} = \frac{20 - 9}{11 - 6} = \frac{11}{5}$$

Solving for E

$$5(64.5 - E) = 11(E - 54.5)$$

$$322.5 - 5E = 11E - 599.5$$

$$16E = 922.0$$

$$E = 57.6$$

By a similar use of the formula, the critical point between the D's and C's proved to be 87.6, and that between the C's and B's was 124.3.

When the contour of either or both curves is irregular in the region of intersection, this formula should not be applied unless and until some reasonable smoothing has been done or until the best-fitting normal frequencies have been substituted for the obtained ones. There may be other instances, depending upon the amount of overlapping, when the similar-triangle principle does not give a good approximation. This can be told by inspection of the plotted curves.

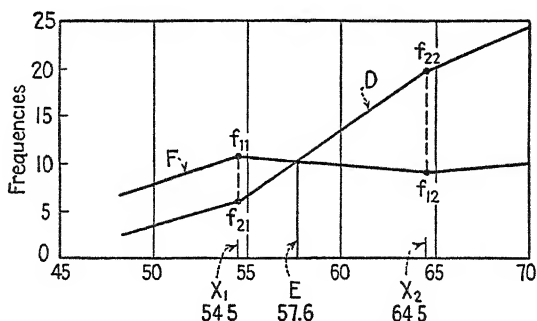


FIG. 29 —An enlarged section taken from Fig. 28 showing the point of intersection of distributions for D and F students

The Accuracy of Prediction of Attributes from Measurements.—

The forecasting efficiency for this kind of prediction is determined in line with that which we found when predicting attributes from other attributes. It is a matter of determining the percentage of correctly predicted cases and the excess of correct prediction. Let us see how well the prediction of F's versus D's would fare in using 57.6 as our critical point. How many D's would be correctly predicted as such, and how many F's correctly predicted as F's? The critical point lies in the interval 50–59 (exactly from 49.5 to 59.5). If it is assumed that the 11 F's within that interval are evenly distributed, then 47.1 F's are above the critical point (wrongly predicted), and 17.9 F's are below this point (correctly predicted). Since we can deal only with whole individuals, we shall round these to 47 and 18, respectively. In the distribution of D's, 179 persons are above a score of 57.6, and only 6 are below that point.

The list of successes and failures is set up as in Table 58. There we see that 79.2 per cent of those with scores above 57.6 were correctly predicted and 75.0 per cent of those with scores below 57.6, which, on the surface, seems fairly high accuracy. But when we realize that

74.0 per cent could have been properly predicted if all 250 cases under consideration had been designated as probable D's, we see that the margin over "guessing" is very small, 6.5 per cent, to be more exact. The success in deciding from known examination score whether a student will receive a mark of D or a mark of F is therefore very limited.

TABLE 58—EFFICIENCY OF PREDICTING A DISCRIMINATION BETWEEN MARKS OF F AND D IN THE ENGLISH COURSE

	Categories		
	F	D	Both
Scores above 57 6	47	179	226
Scores below 57 6	18	6	24
Total	65	185	250

Score range	Prediction	Number correct	Per cent correct
Above 57 6	D	179	79 2
Below 57 6	F	18	75 0
Total		197	78 8
Correct without knowledge	..	185	74 0
Excess with knowledge		12	6 5

It can be seen, incidentally, from Fig. 28, that a discrimination between F and C is also subject to many errors, and even a discrimination between F and B. We should say that all students with scores above about 98 will be B students in this English class and that those below that point will be F's rather than B's, but there would still be some students wrongly predicted. It seems that about the only discrimination of which we could be practically certain is that between F and A, between which, even so, there is overlapping of two or three cases (see Table 57). The forecasting efficiency as between these more remote categories has not been determined. But to show how such efficiency stands up for other neighboring categories, it may be said that the D-C discrimination has a forecasting efficiency of 9.7 per cent in excess of chance, and the C-B discrimination has an excess of 11.2 per cent, both of which are better than that for the F-D discrimination but still nothing to be elated about.

PREDICTING MEASUREMENTS FROM ATTRIBUTES

The Principle of Least Squares.—What would be the most accurate prediction of the weight of a sixteen-year-old youth? By “most accurate,” we mean a weight that, if chosen to predict each sixteen-year-old selected at random from a certain population, would be closer to the facts than any other estimate. To state the matter the other way round, we want a predicted weight that would give us the smallest average discrepancy from the actual weights. For every person, we should find the difference between his actual weight and our prediction in order to obtain the single discrepancy.

Statisticians have good reason to deal here in terms of the *squares* of the discrepancies rather than in terms of the discrepancies themselves. They demand a predicted measurement from which the sum of the squared discrepancies is a minimum. The prediction that will satisfy this requirement has been proved to be the mean of the distribution. In choosing the mean as our prediction, we are following the so-called *principle of least squares*. Whereas in predicting attributes we chose the *mode* of a distribution as the best indicator that would give us the smallest *percentage* of error of placement of cases, in predicting measurements, we choose the *mean* as the indicator, which gives us the smallest set of *squared deviations* from the predicted values.

Predictions Apply to Selected Populations.—So, in answering the question with which we started this discussion, the best prediction of a sixteen-year-old, any better knowledge being lacking, is the mean of the population of which he is a member. If we wanted this to cover *all* sixteen-year-olds, we should see to it that our distribution from which we derive our mean is made up of a large sample in which both sexes, all races, and all socioeconomic and geographic groups are proportionately represented. We might, however, confine the question to American sixteen-year-olds. We might further confine it to high-school youths in one American city or, even further, to one particular high school. Whatever our restriction in population, the predicted weight will apply only to that kind of population, in fact, strictly speaking, it will apply only to the measured sample. Whenever we extend our predictions to samples beyond our known population, we always do so at the risk of enlarging errors of prediction.

Errors of Prediction Measured by the Standard Deviation.—In a certain high school in a certain American city, a random sample of 51 sixteen-year-olds had weights distributed as shown in Fig. 30. For the sake of an illustration, we shall adopt the sixteen-year-olds in this

high school as our population. What we say concerning predictions within this group will hold by analogy to larger, more inclusive populations. The mean of the 51 students is 61.9 kg, and the standard deviation is 13.2. If now the 51 students were listed in alphabetical order and without seeing them we used merely the knowledge of the mean and σ , we should most nearly predict the actual weights if we wrote after each student's name "61.9 kg." The odds are about 2 to 1, as the interpretation of the *SD* goes, that our errors would be no greater than 13.2 kg. either way from the predicted weight. The *SD* of 13.2 kg may therefore be taken to measure our margin of error in

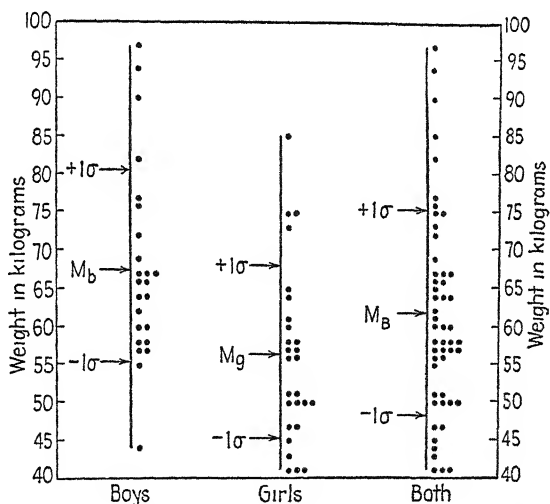


FIG 30 —Distributions of sixteen-year-old high-school boys and girls for weight in kilograms Each dot represents one individual

predicting single cases within the sample, when prediction is based only upon knowledge of the mean.

Any other prediction we might make for all the individuals would yield a larger margin of error, according to the principle of least squares. We should not be very proud of our accuracy of prediction in this instance, and for practical purposes of making decisions for individuals where their weights are important factors, we should be seriously in error in many cases. But we could do less well in predicting the individuals' weights if we did not even possess the knowledge of their mean. Even if we knew the mean of sixteen-year-olds in general and used that as our predictive value, we should do worse than we did, unless the mean of this small population coincides with that of all

sixteen-year-olds. In other words, by knowing one attribute of our population—a group in one American high school—and the mean that goes with that attribute, we reduce the error of prediction to some extent.

Predicting Weight from Knowledge of Sex.—Of the 51 cases in the population of sixteen-year-olds, 24 were boys and 27 were girls. Will it help to predict more accurately if we know each individual's sex? It should, since there is a sex difference in weights. Though many girls are heavier than many boys, the averages are distinctly apart—67.8 for the boys and 56.6 for the girls. Using the attribute of sex to contribute toward the prediction of individual cases and following the principle of least squares, for each boy who came along we should predict his weight to be 67.8 kg, and for each girl, the prediction would be 56.6 kg.

How much will predictions now be improved? The margin of error of predictions for boys is given by the *SD* of their distribution, which is 12.6 kg., and the margin of error for the girls is given by an *SD* of 11.3. From this information, we see that both boys' and girls' weights are more accurately predicted than before (when the margin of error was 13.2) and that the girls' predicted weights are more free from error than are the boys'.

As a matter of consistency with previous procedures, let us ask what the percentage of reduction in error of prediction is. For the boys, the change of .6 in the *SD* is 4.5 per cent, and for the girls, the change in *SD* is 1.9, or 14.4 per cent.

The Standard Error of Estimate.—There is a way of summarizing the margin of error for all cases combined. This requires the computation of a *standard error of estimate*. It is a kind of summary of all the squared discrepancies of actual measurements from the predicted measurements. In terms of a formula, the standard error of estimate is

$$\sigma_{yx} = \sqrt{\frac{\sum(Y - Y')^2}{N}} \quad (51)$$

where Y = measured value of a case we are trying to predict.

Y' = predicted value for the case.

N = total number of cases predicted.

The subscript in σ_{yx} tells us that we are predicting variable Y from variable X . In the illustrative problem, Y is the variable of weight, and X is the variable of sex difference. The sum of the discrepancies squared is 7,287.91; so

$$\sigma_{yx}^2 = \frac{7,287.91}{51} = 142.90$$

and so

$$\sigma_{yz} = 11.9$$

The standard error of the estimate, in predicting weight on the basis of knowledge of sex, is 11.9. Using only the knowledge that this is a particular group of sixteen-year-olds with a mean of 61.9, the error of estimate was given by a standard deviation of 13.2. The margin of error using the information supplied by sex difference is 90.2 per cent as large as that without using this information. The reduction in size of error of prediction is 9.8 per cent, which is rather small but represents some gain.

In computing the standard error of estimate in this kind of problem, it is probably more natural to do so by finding the *SD*'s of the two part distributions separately and then combining them. They cannot be combined directly by simple addition or averaging. It is the squared deviations in the two groups that must be combined. The sum of the squared deviations in each distribution can be found by the formula

$$\Sigma x_a^2 = N_a \sigma_a^2 \quad (52)$$

where Σx_a^2 = sum of the squared discrepancies between prediction and fact (or between measurements and the mean) in distribution *A* (one of the attribute distributions).

N_a = number of cases in distribution *A*.

σ_a = standard deviation of distribution *A*.

When these sums of squared deviations are obtained from all component distributions (distributions *A*, *B*, and *C*, etc.), they may be combined by simple addition to give $\Sigma(Y - Y')^2$. In other words

$$\Sigma(Y - Y')^2 = \Sigma N_k \sigma_k^2 \quad (53)$$

where N_k = number of cases in any component distribution (distributions *A*, *B*, *C*, etc., in turn).

σ_k = standard deviation of the same distribution.

The work of computing $\Sigma(Y - Y')^2$ for the problem on weights of sixteen-year-olds may be summarized as follows:

Distribution	<i>N</i>	σ	σ^2	$N\sigma^2$
Boys.	24	12.65	160.02	3,840.48
Girls	27	11.30	127.69	3,447.63
				7,288.11
				$\Sigma(Y - Y')^2$

From here the computation of σ_{yz} is exactly the same as previously demonstrated.

Other Predictive Indices May Be Introduced.—It should be added that other attributes may be brought into the predictive picture. For example, if different glandular constitution has a definite bearing on body weight, for example, thyroid functioning, we could subdivide each sex group into two or three categories as to glandular condition. The mean of each new subgroup would then become the prediction for members of that group. The deviations of actual weights from these means would be smaller and the new standard error of estimate would be reduced in size.

If we were successful in singling out all the significant factors correlated with weight and could predict from all of them at the same time, theoretically we could reduce errors of prediction to approximately zero. We can probably never know what all the significant factors are from which weight can be determined, and if we did it might be impossible to assign all the attributes to each individual. We are here speaking of the hypothetical limiting case. Any improvement in predictions approaches that limit. From a practical standpoint, it is always a question of whether the trouble of uncovering and using new descriptive attributes is justified by the gains in predictive accuracy that result.

PREDICTING MEASUREMENTS FROM OTHER MEASUREMENTS

When both known and predicted variables are measured on linear scales and there is some relation between them so that predictions are possible, we have a much more complicated problem, which we shall merely introduce in this chapter. A complete treatment of it involves correlation methods and regression equations, which are treated in the next chapter.

The Correlation Diagram.—Our illustration of this kind of problem consists of two achievement examinations in a course on educational measurements. In Table 59, we have the two distributions grouped in class intervals and the measurements in each class interval broken down to form a distribution of its own in the other test. The class intervals for test *X* are listed along the top of Table 59 and the class intervals for test *Y* are listed along the left margin. The 26 individuals who made scores with in the interval 75–79 in test *X* made scores in test *Y* that distribute themselves all the way from the interval 100–104 to the interval 130–134. The 22 cases that made scores in the

interval 115-119 for test *Y* distribute themselves all the way from the interval 70-74 to the interval 90-94 for test *X*. In a similar manner,

TABLE 59.—PREDICTING SCORES IN ONE TEST FROM KNOWN SCORES IN ANOTHER TEST

Test <i>Y</i>	Test <i>X</i>								f_y	M_{row}	σ_{row}
	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99			
135-139								1	1	97 0	0 00
130-134				1	1	0	1		3	83 7	6 61
125-129				1	0	2	1		4	85 8	5 45
120-124			1	4	4	6	2		17	83 2	5 67
115-119			7	5	7	2	1		22	78 6	5 72
110-114	1	4	2	9	4	2			22	75 9	6 56
105-109	1	1	2	5	1				10	74 0	5 56
100-104	1	3	0	1	1				6	70 3	6 87
95- 99		2							2	67 0	0 00
f_x	3	10	12	26	18	12	5	1	87 = <i>N</i>		
M_c	107 0	105 5	114 9	114 5	116 4	120 3	124 0	132 0			
σ_c	4 08	5 52	4 31	6 83	6 43	4 71	5 10	0 00			

one can see from the table, which we call a *correlation diagram*, how the individuals falling in any interval in the one test distribute themselves in the other test.

Prediction of *Y* from *X*.—

As usual, we have here a double prediction problem; the prediction of a score in *Y* from a known score in *X*, and vice versa. Let us consider the prediction of *Y* from *X* first. For the individuals in any class interval in test *X*, the best prediction is the mean of the *Y*-distribution in that column, in other words, the mean of the column, M_c . For each column of Table 59, its mean is listed in the next to the last row. For the first column, M_c is 107.0. Any

person receiving a score from 60 to 64 inclusive in test *X* will most probably earn a score of 107.0 in test *Y*. The other means of the

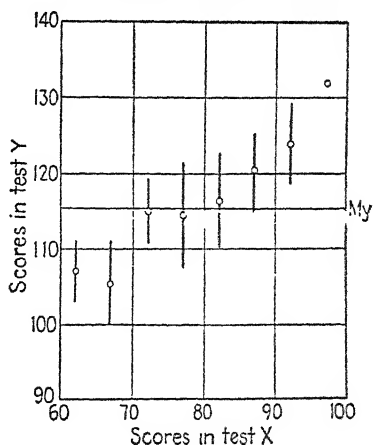


FIG. 31.—A chart showing the most probable score in Test *Y* corresponding to each midpoint score in Test *X*, also the range between minus and plus one standard deviation

columns are similarly interpreted. It will be noticed that there is a general upward trend in the M_c 's as we go up the scale in test X , though there are two inversions. In view of the small numbers of cases upon which these means are based, some inversions are not surprising.

The margin of error in predicting Y from X in each column is the standard deviation of that column. The σ_c 's are listed in the last row of Table 59. They remain fairly constant, except for the σ of zero, which is based upon the exceptional distribution of one case and so can be ignored. The entire picture of predictions and their margins of errors within columns is shown graphically in Fig. 31. The circlets show the positions of the column means, and the vertical lines running through them extend from $-1\sigma_c$ to $+1\sigma_c$. In each column, we expect two-thirds of the observed scores to lie within the limits of these lines.

Standard Error of Estimate.—In order to obtain a single indicator of the goodness of the predictions of Y -scores from X -scores, we may compute a standard error of estimate as we did before when predicting measurements from attributes. The work is best organized as in Table 59a. For every column, we list first N_c , the number of cases in that

TABLE 59a.—COMPUTATION OF THE STANDARD ERROR OF ESTIMATE OF Y -SCORES FROM X -SCORES

N_c	σ_c	$N_c\sigma_c^2$
3	16.67	50.00
10	30.45	304.50
12	18.58	223.00
26	46.63	1,212.50
18	41.36	744.48
12	22.22	226.68
5	26.00	130.00
1	0.00	0.00
		2,931.16 = $\Sigma(Y - Y')^2$

column. Second, we list σ_c^2 , the squared SD of the distribution in that column. Next we find the product of these two values for that column. The sum of these products for all columns yields $\Sigma(Y - Y')^2$, which we need for computing $\sigma_{y \cdot x}$. This sum is 2,931.16. From here on the work follows formula (51).

$$\sigma_{y \cdot x}^2 = \frac{2,931.16}{87} = 33.6915$$

and so

$$\sigma_{yx} = 5.80$$

The *SD* of the entire distribution of *Y*-scores is 7.85, which gives us a reduction in variability of 2.05, or 26.1 per cent, a marked improvement in prediction, as such tests go. We may say that the forecasting efficiency for predicting *Y*-score from *X*-score as we did is 26.1 per cent.

Predicting *X* from *Y*.—The predictions of *X* from *Y* are listed in Table 59 under *M_{row}* in the next to the last column. The most probable *X*-score for any interval of *Y*-scores is the mean of the row. The margin of error of the predictions is given in each case by σ_{row} , and these appear in the last column of Table 59. To complete the picture of these predictions and their *SD*'s, Fig. 32 is presented. The standard error of estimate of the *X*-scores, σ_{xy} (note the order of *x* and *y* in the subscript), is equal to 5.93. Since the total *SD* of the *X*-scores is 7.55, the reduction in error of prediction is 1.62, which is 21.5 per cent. The forecasting efficiency in predicting *X* from *Y* is in this problem somewhat lower than the forecasting efficiency (26.1 per cent) in predicting *Y* from *X*.

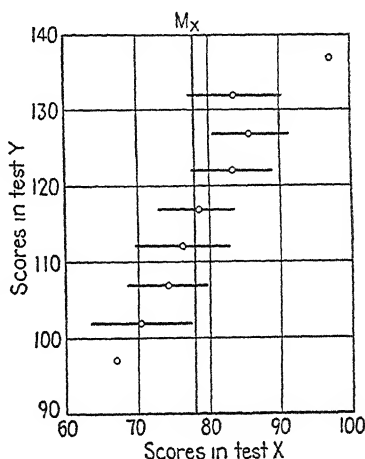


FIG. 32.—A chart showing the most probable score in Test *X* for each interval of scores in Test *Y*; also the range of scores in Test *X* within each row.

Regression Lines.—In practice, we should rarely make predictions of measurements from other measurements as we have illustrated here. There are simpler and more general procedures for doing so. It will be seen in Figs. 31 and 32 that the trend of the means of the columns and of the rows in each case approaches a straight line. We may assume that there is in reality a straight-line (or rectilinear) relationship between *Y* and *X* and that the unreliable means that we obtained deviate from it by chance fluctuations. Correlation methods to be explained in the next chapter enable us to find the line that will come nearest to the means of the columns (or rows). Our predictions will then lie along those lines. There will be one *regression line* for the regression of *Y* on *X* and another line for the regression of *X* on *Y*. This follows from the two-way prediction problem that we have.

If the trend in a series of column means should be curved rather than obviously straight, then we should either assume some known type of mathematical curve to fit the points, along which to make predictions or we should compute predictions as we have done in this chapter. In this chapter, we have used the standard error of estimate and the percentage of reduction of errors as indices of goodness of prediction; in the next chapter, we shall see the importance of the coefficient of correlation as an indicator of the same thing and of other things about the relationship between two or more variables.

Exercises

1. In Data *X*, make predictions in both directions, and determine the percentages of correct predictions, the number of correct predictions without knowledge of another attribute, and the percentage of forecasting efficiency of each prediction. Discuss the usefulness of these predictions

DATA *X*—RELATIONSHIP BETWEEN FAILING IN COLLEGE AND BEING ABOVE OR BELOW THE MEDIAN IN HIGH-SCHOOL GRADUATING CLASS

Status in high-school class	Failing in one or more courses	No failures in first semester	Total
Above the median	37	340	377
Below the median	49	71	120
Total	86	411	497

2. In Data *Y*, make predictions of whether a student will report "Yes," "?" or "No" to the question about talking when he makes similar responses to the question about walking in his sleep, and vice versa. How accurate are these predictions?

DATA *Y*.—RELATIONSHIP BETWEEN WALKING IN ONE'S SLEEP AND TALKING IN ONE'S SLEEP AS REPORTED BY 1,781 STUDENTS*

Talk in your sleep?	Walk in your sleep?			
	Yes	?	No	Total
Yes	88	9	400	497
?	3	14	194	211
No	7	3	1,069	1,079
Total	98	26	1,663	1,787

* Jenness, A. F., and Jorgensen, A. P. Ratings of vividness of imagery in the waking state compared with reports of somnambulism. *Amer. J. Psychol.*, 1941, 54, 253-259. Reproduced with the permission of the editor of *Amer. J. Psychol.*

3. For Data *C* (page 27), find the critical division point between the two chemistry classes. Set up a table summarizing your predictions of group from knowledge of scores, and find the forecasting efficiency.

4. For Data *C*, what is the best predicted score for each of the two groups? What is the margin of error of prediction for each group? For the two combined? What is the index of forecasting efficiency?

5. For Data *Z*, find the best prediction of score in the opposites test for each midpoint score in the mixed-sentences test. Find the margin of error for each prediction and for the predictions taken as a whole.

DATA *Z*—A SCATTER DIAGRAM FOR TWO MENTAL TESTS

<i>Y</i> (Opposites test in Army Alpha)	<i>X</i> (Mixed-sentences test in Army Alpha)							
	0-2	3-5	6-8	9-11	12-14	15-17	18-20	21-23
36-38								1
33-35							1	2
30-32				1	1	3	7	2
27-29						4	5	2
24-26			1	3	3	2	4	4
21-23			1		6	1	5	2
18-20		1	2	1	9	5	4	
15-17	2	1	2	2	2	2	1	
12-14	1	2	0	2	2	1		
9-11	3	1	2	1	2			
6-8				1				
<i>f_x</i>	6	5	8	11	25	18	27	13

CHAPTER XI

CORRELATION METHODS

No single statistical procedure has opened up so many new avenues of discovery in psychology and education as that of correlation. This is understandable when we remember that scientific progress depends upon finding out what things are co-related and what things are not. A *coefficient of correlation* is a single number that tells us to what extent two things are related; to what extent variations in the one go with variations in the other. Without the knowledge of how one thing varies with another, we should find predictions impossible. And wherever causal relationships are involved, without knowledge of co-variation, we should be unable to control one thing by manipulating another.

For example, when we know that the higher a girl's score in a clerical-aptitude test the higher the average performance she is likely to exhibit after training, we can thereafter use scores on this test to predict level of proficiency. We say that there is a high positive correlation between aptitude-test score and clerical success. We discover this fact by finding a coefficient of correlation between scores of a number of girls and measures of clerical performance later for the very same girls. We can never compute a coefficient of correlation on one person alone, nor can we compute it without having made two sets of measurements on the same individuals. In this instance, if we consider that the aptitude test has measured individual differences in some quality or qualities that lead to success, *i.e.*, in the sense of a cause of clerical success, then we can not only predict future success for individuals but also promote high general efficiency in any group of clerks by selecting those with high scores. Thus are studies leading to prediction and control of human affairs promoted because correlation techniques are available. Without some device for checking up like this on a test, we have only vague notions concerning its effectiveness, unless, indeed, its effectiveness is so obvious to direct observations as to require no inspection by correlation methods, which is highly unlikely.

Some Examples of Correlation between Two Variables.—The coefficient of correlation is one of those summarizing numbers, like a

mean or a standard deviation, which, though it is a single number, tells a story. It can vary from a value of $+1.00$, which means perfect positive correlation, through zero, which means complete independence or no correlation whatever, on down to -1.00 , which means perfect negative correlation.

A Case of Perfect, Positive Correlation—Figure 33 illustrates an instance of perfect positive correlation. It is a fictitious case, for such exact agreement between two things is rarely or never experienced, certainly not in psychology or education. Here we have assumed two tests, X and Y . Ten individuals have received scores in the two tests. The pairs of scores are as follows:

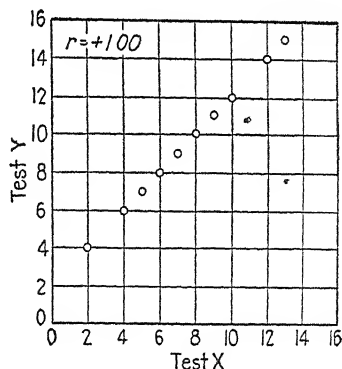


FIG 33.—A simple correlation chart showing the relationship between X and Y scores when the correlation is $+1.00$

Individual.	A	B	C	D	E	F	G	H	I	J
Score in test X	2	4	5	6	7	8	9	10	12	13
Score in test Y	4	6	7	8	9	10	11	12	14	15

Looking down the rows of scores, each pair made by one individual, we readily conclude that each person's score in Y is two points higher than his score in X . In terms of a simple equation, $Y = X + 2$. There are *no exceptions*, which makes the correlation perfect.

To take another instance:

Individual	A	B	C	D	E	F	G	H	I	J
Score in test P	1	3	4	5	7	8	9	11	12	15
Score in test Q	2	6	8	10	14	16	18	22	24	30

In this situation, each person's score in Q is two times that in P , again without exception; there is perfect agreement, and the coefficient of correlation would be $+1.00$. The equation for predicting Q from P is $Q = 2P$.

A Case of High Positive Correlation.—In Fig. 34, we have illustrated a case of correlation that is positive but less than $+1.00$. The graphic

picture of the individuals shows that, in general, a person who is high in test X is also high in test Y , and one who is low in X is also likely to be low in Y . The actual scores for these 10 people are listed in the first two columns of Table 60. It will be seen that although the individuals are arranged in rank order for scores in X , there are some deviations from this rank order when we inspect their scores in Y . The coefficient of correlation by computation is equal to $+0.76$. We

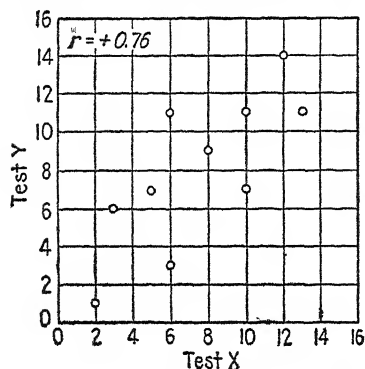


FIG. 34.—A correlation chart illustrating the situation when the coefficient is $+0.76$.

shall soon see how this was obtained but first simply note by comparison of Figs. 33 and 34 how the individuals are scattered in the diagrams. In Fig. 33, they line up in perfect file from lowest to highest. In Fig. 34, they tend to fan out to diverge from a strict line-up, but a definite trend of relationship can be observed. The amount of spreading in Fig. 34, as compared with that in Fig. 33 (in which it is, of course, none), illustrates the difference between correlations of $+1.00$ and $+0.76$.

A Case of Low Positive Correlation.—A third instance is shown in Fig. 35, in which the spreading effect to which our attention was called before is even greater. The coefficient of correlation is here $+0.14$; in other words, close to zero. This being true, a person with high score

TABLE 60—CORRELATION BETWEEN TWO SETS OF MEASUREMENTS OF THE SAME INDIVIDUALS, UNGROUPED DATA; PRODUCT-MOMENT COEFFICIENT OF CORRELATION

X	Y	x	y	x^2	y^2	xy
13	11	+5 5	+3	30 25	9	+16 5
12	14	+4 5	+6	20 25	36	+27 0
10	11	+2 5	+3	6 25	9	+ 7.5
10	7	+2 5	-1	6 25	1	- 2.5
8	9	+0 5	+1	0 25	1	+ 0 5
6	11	-1 5	+3	2 25	9	- 4.5
6	3	-1 5	-5	2 25	25	+ 7.5
5	7	-2 5	-1	6 25	1	+ 2.5
3	6	-4 5	-2	20 25	4	+ 9 0
2	1	-5 5	-7	30 25	49	+38.5
Sums 75	80	0 0	0	124.50	144	102 0
Means 7.5	8 0			Σx^2	Σy^2	Σxy

$$\sigma_x = \sqrt{\frac{124.50}{10}} = \sqrt{12.450} = 3.53$$

$$\sigma_y = \sqrt{\frac{144}{10}} = \sqrt{14.4} = 3.79$$

$$r_{xy} = \frac{\Sigma xy}{N\sigma_x\sigma_y} = \frac{102.0}{(10)(3.53)(3.79)} = \frac{102.0}{133.787} = +.76$$

An alternative solution without computing the σ 's:

$$r_{xy} = \frac{\Sigma xy}{\sqrt{(\Sigma x^2)(\Sigma y^2)}} = \frac{102.0}{\sqrt{(124.5)(144)}} = \frac{102.0}{\sqrt{17,928.0}} = \frac{102.0}{133.9} = +.76$$

in X is likely to be almost anywhere, within the total range, in terms of his Y -score. The three highest people in X , with scores of 10, 12, and 13, scatter all the way from 3 to 11 in test Y . The three lowest people in test X , with scores of 1, 3, and 4, scatter all the way from 2 to 9 in test Y . Although there is a trace of relationship between X -scores and Y -scores, it is very weak. The actual scores may be compared in Table 62.

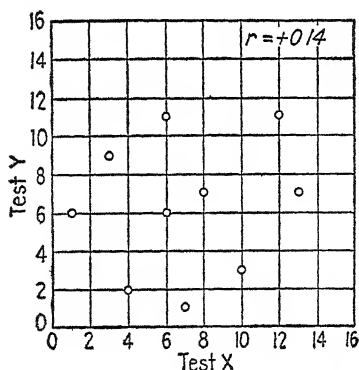


FIG. 35.—A correlation chart when the correlation is only +.14.

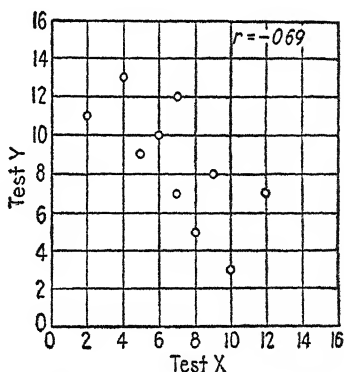


FIG. 36.—A correlation chart with a correlation of $-.69$.

A Case of High Negative Correlation.—The situation that obtains when there is a negative correlation is shown in Fig. 36. Here the coefficient is $-.69$. Compare this diagram with that in Fig. 34, and it will be apparent that the trend of the points is along the other diagonal now, from upper left to lower right. This illustrates the fact that persons making high scores in X are likely to make low scores in Y , and persons making low scores in X are likely to make high scores in Y . This inverse order of relationship is also apparent in the actual scores in the first two columns of Table 61. The numerical size of the

TABLE 61.—A NEGATIVE CORRELATION IN UNGROUPED DATA BY THE PRODUCT-MOMENT METHOD

<i>X</i>	<i>Y</i>	<i>x</i>	<i>y</i>	<i>x</i> ²	<i>y</i> ²	<i>xy</i>
12	7	+5	-1 5	25	2 25	- 7 5
10	3	+3	-5 5	9	30 25	-16 5
9	8	+2	-0 5	4	25	- 1 0
8	5	+1	-3 5	1	12 25	- 3 5
7	7	0	-1 5	0	2 25	0 0
7	12	0	+3 5	0	12 25	0 0
6	10	-1	+1 5	1	2 25	- 1 5
5	9	-2	+0 5	4	.25	- 1 0
4	13	-3	+4 5	9	20 25	-13 5
2	11	-5	+2 5	25	6 25	-12 5
Sums 70	85	0	0 0	78	88 50	-57 0
Mean 7 0	8 5			Σx^2	Σy^2	Σxy

$$\sigma_x = \sqrt{\frac{78}{10}} = \sqrt{7.8} = 2.79$$

$$\sigma_y = \sqrt{\frac{88.5}{10}} = \sqrt{8.85} = 2.97$$

$$r_{xy} = \frac{-57.0}{(10)(2.79)(2.97)} = \frac{-57.0}{82.863} = -.69$$

coefficient (.69) is nearly the same as for the correlation in Fig. 34 (.76). It will be seen that the width of scatter of the points is about the same in the two cases. A perfect negative correlation would be pictured as a line of dots like that in Fig. 33 but it would slant downward instead of upward from left to right. The algebraic sign of the coefficient of correlation therefore merely has to do with the *direction* of the relationship between two things, whether direct or inverse, and the size of the coefficient (distance from zero) has to do with the *strength* or *closeness* of the relationship.

HOW TO COMPUTE A COEFFICIENT OF CORRELATION

The Product-moment Coefficient of Correlation.—The standard kind of coefficient of correlation and the one most commonly computed is Pearson's product-moment coefficient. The basic formula is

$$r_{xy} = \frac{\Sigma xy}{N\sigma_x\sigma_y} \quad (54)$$

where r_{xy} = correlation between X and Y .

x = deviation of any person from the mean in test X .

y = deviation of the same person from the mean in test Y .

Σxy = sum of all the individual products, each person's x deviation times his y deviation.

N = number of persons measured.

σ_x and σ_y = standard deviations of the distributions of X - and Y -scores. The steps necessary are illustrated in Table 60. They will be enumerated here:

- Step 1. List in parallel columns the individuals' X and Y scores, making sure that each person's two scores are together.
- Step 2. Determine the two means M_x and M_y . In Table 60, these are 7.5 and 8.0, respectively.
- Step 3. Determine for every person his two deviations x and y . Check them by finding algebraic sums, which should be zero.
- Step 4. Square all the deviations, and list in two columns. This is for the purpose of computing σ_x and σ_y .
- Step 5. Sum the squares of the deviations to obtain Σx^2 and Σy^2 .
- Step 6. From these values compute σ_x and σ_y , saving one more than the number of significant digits.
- Step 7. For every person, find his xy product (last column of Table 60). Sum these for Σxy .
- Step 8. You are now ready for formula (54). In the illustrative problem, the arithmetic is given following Table 60 (page 201).

A Shorter Solution.—There is an alternative and shorter route that omits the computation of σ_x and σ_y , should they not be needed for any other purpose. The formula is

$$r_{xy} = \frac{\Sigma xy}{\sqrt{(\Sigma x^2)(\Sigma y^2)}} \quad (55)$$

The solution with this formula is also given on page 201, and it leads to the same coefficient. In both cases, two significant digits have been saved in r , though three might have been legitimately saved, for the reason that for so small a number of cases the sampling error in r is so relatively large that more than two digits would be rather deceiving as to accuracy. When N is large—100 or more—three-place accuracy in r may more properly be recognized.

Computing a Negative Coefficient.—As another example of the computation of r , when the correlation is *negative*, Table 61 is sub-

mitted. The operations are just the same, step by step. The only thing new is the care that must be taken with algebraic signs.

Computing r from Original Measurements.—In both examples thus far, we have been dealing with a small number of observations and ungrouped data. When the data are more numerous, we resort to grouping into class intervals; but first let us see another procedure with ungrouped data, which does not require the use of deviations. It works entirely from original scores. When raw scores are small numbers or when a good calculating machine is available, this is the best procedure. The formula looks forbidding but is really easy to apply:

$$r_{xy} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}} \quad (56)$$

where X and Y = original scores in tests X and Y . Other symbols tell what is done with them. We follow the steps that are illustrated in Table 62.

TABLE 62.—CORRELATION OF UNGROUPED DATA COMPUTED FROM THE ORIGINAL MEASUREMENTS

X	Y	X^2	Y^2	XY
13	7	169	49	91
12	11	144	121	132
10	3	100	9	30
8	7	64	49	56
7	2	49	4	14
6	12	36	144	72
6	6	36	36	36
4	2	16	4	8
3	9	9	81	27
1	6	1	36	6
Sums 70	65	624	533	472
ΣX	ΣY	ΣX^2	ΣY^2	ΣXY

$$\begin{aligned}
 r_{xy}^2 &= \frac{[N\Sigma XY - (\Sigma X)(\Sigma Y)]^2}{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]} \\
 &= \frac{(4,720 - 4,550)^2}{(6,240 - 4,900)(5,330 - 4,225)} \\
 &= \frac{(170)^2}{(1,340)(1,105)} \\
 &= \frac{28,900}{1,480,700} \\
 &= .019518 \\
 r_{xy} &= \sqrt{.019518} \\
 &= +.14
 \end{aligned}$$

- Step 1. Square all X and Y measurements.
- Step 2. Find the XY product for every person.
- Step 3. Sum the X 's, the Y 's, the X^2 's, the Y^2 's, and the XY 's.
- Step 4. Apply formula (56).

The writer has found it more convenient, particularly when machine work can be done, to compute r^2_{xy} first by the formula

$$r^2_{xy} = \frac{[N\Sigma XY - (\Sigma X)(\Sigma Y)]^2}{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]} \quad (57)$$

and then finally extract the square root to find r_{xy} , as shown just below Table 62.

Preparing a Scatter Diagram.—When N is large, even when N is moderate in size, and when no calculating machine is available, the customary procedure is to group data in both X and Y and to form a scatter diagram or correlation diagram. The choice of size of class interval and limits of intervals follows much the same rules as were given in Ch. II. For the sake of a clearer illustration of the procedure, a smaller number of classes will be employed in the problem now to be described. The data were scores earned by a class in educational measurements in two objectively scored examinations, one of which stressed statistical methods and the other of which stressed tests and measurements. The same data were used to illustrate the methods of predicting measurements from other measurements in the preceding chapter. We use them again not only to show how a coefficient of correlation is computed in grouped data but also to show how predictions can be made by means of regression equations, which are related to the coefficient of correlation.

In setting up a double grouping of data, a table is prepared with columns and rows—columns for the dispersions of Y -scores within each class interval for the X -scale, and rows for the dispersions of X -scores within each class interval for the Y -scale. Along the top of the table (see Table 63) are listed the score limits for the class intervals in test X . Along the left-hand margin are listed the score limits for the class intervals in test Y . We make one tally mark for each individual's X - and Y -scores. For example, if one individual had a score of 83 in test X and a score of 121 in test Y , we place a tally mark for him in the *cell* of the diagram at the intersection of the column for interval 80–84 in X and the row for interval 120–124 in Y . All other individuals are similarly located in their proper cells.

When the tallying is completed, we write the number of cases, or the *cell frequency*, in each of the cells. Next we sum the cell frequencies

in the rows separately, recording each frequency in the last column under the heading f_y . When this column is filled, we have the total frequency distribution for test Y . We also sum the cell frequencies in all the columns, writing them in the bottom row with its heading f_x . When completed, this row gives us the total frequency distribution for test X . We can check the sum of the cell frequencies by adding up the last row and last column. Their sums should, of course, both equal N , in this case, 87. The check does not, however, guarantee correct tallying. This can be checked partly when we correlate either test with another one and compare total frequency distributions or

TABLE 63—A SCATTER DIAGRAM OF THE SCORES IN TWO ACHIEVEMENT TESTS
X. Scores in First Achievement Test

X: Scores in First Achievement Test									
	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99	f_y
135-139								/ 1	1
130-134			/ 1	/ 1			/ 1		3
125-129			/ 1			// 2	/ 1		4
120-124			/ 1	/// 4	/// 4	/// 6	// 2		17
115-119			/// 7	/// 5	/// 7	// 2	/ 1		22
110-114	/ 1	/// 4	// 2	/// 9	/// 4	// 2			22
105-109	/ 1	/ 1	// 2	/// 5	/ 1				10
100-104	/ 1	/// 3		/ 1	/ 1				6
95-99		// 2							2
f_x	3	10	12	26	18	12	5	1	87

when we have knowledge of the correct frequency distribution of Y or of X from any other source. There are times when it is wise to do the entire tallying two times and to compare all cell frequencies in the two attempts. It is very easy to place a tally mark in the wrong cell.

Computing the Pearson r from a Scatter Diagram.—When the product-moment r is computed from a scatter diagram, the formula becomes

$$r_{xy} = \frac{\frac{\sum x'y'}{N} - (c'_x c'_y)}{(\sigma'_x)(\sigma'_y)} \quad (58)$$

where x' and y' = deviations from the guessed mean in terms of the class interval as the unit.

c'_x and c'_y = corrections in X and Y .

σ'_x and σ'_y = standard deviations in X and Y in terms of the class interval as the unit.

The details of application of this equation will now be explained and illustrated

Determining the Corrections and Standard Deviations.—The procedure for calculating the corrections and standard deviations for both X and Y separately are no different than was previously described

TABLE 64—SCATTER DIAGRAM FOR COMPUTING A PEARSON r
X Examination in Statistics

	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99	f_y	y'	fy'	fy'^2	$\Sigma x'y'$
													$\begin{smallmatrix} + \\ - \end{smallmatrix}$
135-139								¹⁶ 1 ₁₆	1	+4	+4	16	16
130-134				1	³ 1 ₃	⁶	³ 1 ₉		3	+3	+9	27	12
125-129				1	²	⁴ 2 ₈	⁶ 1 ₆		4	+2	+8	16	14
120-124			⁻¹ 1 ₋₁	4	¹ 4 ₄	² 6 ₁₂	³ 2 ₆		17	+1	+17	17	22
115-119			7	5	7	2	1		22	0	0	0	0
110-114	³ 1 ₃	² 4 ₈	¹ 2 ₂	9	⁻¹ 4 ₋₄	⁻² 2 ₋₄			22	-1	-22	22	13
105-109	⁶ 1 ₆	⁴ 1 ₄	² 2 ₁	5	⁻² 1 ₋				10	-2	-20	40	14
100-104	⁹ 1 ₉	⁶ 3 ₁₈		1	⁻³ 1 ₋₃				6	-3	-18	54	27
95-99		⁸ 2 ₁₆							2	-4	-8	32	16
f_x	3	10	12	26	18	12	5						
x'	-3	-2	-1	0	+1	+2	+3	+4			$\Sigma fy'$	$\Sigma fy'^2$	
fx'	-9	-20	-12	0	+18	+24	+15	+4	+20		$= \Sigma fx'$		
fx'^2	27	40	12	0	18	48	45	16	206		$= \Sigma fx'^2$		
$\Sigma x'y'$	18	46	6	0	7	20	21	16	134		$\Sigma x'y'$	+120	
			1	0	9	4			-14				

(see pages 54*f*). From Table 64 we have the necessary information, which is used as follows:

$$c'_x = \frac{\Sigma fx'}{N} = \frac{20}{87} = .230$$

$$c'_y = \frac{\Sigma fy'}{N} = \frac{-30}{87} = -.345$$

$$\sigma'_x = \sqrt{\frac{\Sigma fx'^2}{N} - (c'_x)^2} = \sqrt{\frac{206}{87} - .0529} = \sqrt{2.3161} = 1.52$$

$$\sigma'_y = \sqrt{\frac{\Sigma fy'^2}{N} - (c'_y)^2} = \sqrt{\frac{224}{87} - .1190} = \sqrt{2.4557} = 1.57$$

Determining the Sum of the Cross Products.—The new process to be mastered here is the calculation of the cross products, or products of the moments, and their sum, in other words, $\Sigma x'y'$. It is best to begin with the idea that every cell has its own $x'y'$ product and to keep that idea in mind. In fact, it is well to determine the $x'y'$ product for every cell in which individuals fall and to write it in as was done in Table

64. The $x'y'$ product for any cell is simply the product of the x' value times the y' value of that cell, close watch being kept of algebraic signs. This matter is easily checked, of course, by making sure that the sign of every $x'y'$ product is positive in the upper right quarter of the chart and also the lower left quarter, but they are all negative in the upper left and lower right quarters. This rule presupposes that the X measurements are increasing from left to right and that the Y measurements are increasing from below upward.

Having given every cell its $x'y'$ value and having recorded it in the upper left-hand corner of the cell, we next note how many individuals have that $x'y'$ value—in other words, the frequency in that cell. We multiply the cell product by the frequency, and in Table 64 these products are recorded with algebraic sign in the lower right-hand corners of the cells. All that remains now is to summate them. We do this both in the columns and in the rows for the sake of checking, for this is an unusually critical number in the correlation formula, and because of the many steps involved in deriving it there are many opportunities for errors. The last two columns in Table 64 are devoted to the sums of $fx'y'$ values in the rows. We keep the sums of the positive products in one of these columns and the sums of the negative products in the other. The last two rows of the table are reserved likewise for summing the positive and negative sums in the columns. Summing everything in the last two columns (also in the last two rows) of the Table gives us $\Sigma x'y'$, and the two estimates should check exactly. For the illustrative problem, the positive sum is 134 and the negative is -14, leaving a net positive sum $\Sigma x'y'$ of 120. We now have everything we need for calculating r . Applying formula (58), we have

$$\begin{aligned} r_{xy} &= \frac{\frac{120}{87} - (.23)(-.345)}{(1.52)(1.57)} \\ &= \frac{1.3793 + .0794}{2.3864} \\ &= \frac{1.4587}{2.3864} \\ &= .61 \end{aligned}$$

The Statistical Reliability of a Coefficient of Correlation.—Like every statistic, the coefficient of correlation is subject to errors of sampling and of measurement. The computed coefficient is derived from a limited sample. The true coefficient that would be obtained if the entire population were measured with perfect measurements we

cannot know. The correlations in successive samples like the one we measured would fluctuate about the true coefficient as their mean, with a standard error of a certain size. If we knew this standard error, we could say how much deviation from the true r any sample of our size would probably exhibit. If we assume that the true coefficient is equal to our obtained one, we can estimate the standard error of sample coefficients about this value by means of the formula

$$\sigma_r = \frac{1 - r^2}{\sqrt{N - 1}} \quad (59)$$

In the correlation problem we have just solved

$$\begin{aligned} \sigma_r &= \frac{1 - .61^2}{\sqrt{86}} \\ &= \frac{1 - .3721}{9.2736} \\ &= \frac{.6279}{9.2736} \\ &= .068 \end{aligned}$$

We may interpret this standard error as in similar instances. We may say that if .61 is the true r , the odds are about 2 to 1 that r 's in samples of 87 cases each will fall between .542 and .678. Or, to frame the interpretation in a more useful manner, we can say that the odds are 2 to 1 that the true r does not lie below .542 or above .678. Allowing margins of $3\sigma_r$ each way from .61, we can say that we are almost certain that the true r lies between .61 plus or minus $20\frac{1}{2}$, or between .406 and .814.

It should be added, however, that r is not like other statistics in one respect; the distribution of samples of r is not always normal or even symmetrical. For small and moderate values of r , when N is not too small, normality is to be depended upon, but when r exceeds .80 or thereabouts (either positive or negative .80), the distributions are skewed. For r 's of large size, therefore, it is of little significance to compute a standard error of r itself. Fisher has met this difficulty by transforming r 's into another statistic known as z (not to be confused with a standard measurement), which is normally distributed. Then, an equation for the SE of z being provided, all the usual tests of significance based upon an SE can be applied, including the significance of differences between two z 's. Because the occasions in which one needs to test significance of differences between r 's are rather rare,

Fisher's use of z is not described here. The description can be found in Fisher's own writings or in advanced texts on statistics.¹

It is frequently asserted that when an r is at least three times its SE , it is a significant correlation. What is meant is that, if the line of reasoning in preceding paragraphs is followed, there would be an extremely small chance for the true r to be zero or less (zero or negative when the obtained r is positive; zero or positive when the obtained r is negative). The chance would be given by the area under the tail of the normal distribution curve beyond a $z = 3.0$; in other words, a probability of .0014, or 14 chances in 10,000.

But there is a better way of determining this kind of significance. And particularly do we want to know this kind of fact about our obtained r when it is rather small and when N is relatively small, and we suspect that therefore there may be a chance that our r could have arisen out of a population in which the two variables in question are really not correlated. We therefore set up the hypothesis that the two variables as measured are not correlated, that the true r equals zero in our population. We ask, then, how likely is an r as large as the obtained one or larger in samples from this population. With a true r of zero, the formula for σ_r becomes

$$\sigma_{r_0} = \frac{1}{\sqrt{N-1}} \quad (60)$$

In our illustrative problem

$$\begin{aligned} \sigma_{r_0} &= \frac{1}{\sqrt{86}} \\ &= \frac{1}{9.2736} \\ &= .108 \end{aligned}$$

Imagine a distribution of r 's with zero at the mean and a σ of .108. Our obtained r is .61, which is 5.65 times this SE . This value is a t ratio and may be interpreted as such. It is obvious that t is so large as to overthrow the hypothesis of no correlation. In other instances, the test of the null hypothesis would not be so decisive—for example, the first problems in correlation mentioned in this chapter.

Even better than computing a t ratio is the practice of consulting Table D in the Appendix to find how large r should be for different levels

¹ See Peters, C. C., and Van Voorhis, W. R., *Statistical methods and their mathematical bases*. New York: McGraw-Hill, 1940. Pp. 155-157, 185-189. Also, Lindquist, E. F., *Statistical analysis in educational research*. Boston: Houghton, 1940. Pp 210ff.

of significance (5 per cent level or 1 per cent level) for varying numbers of cases. The table is entered with the number of degrees of freedom, which, in a simple correlation problem, is $N - 2$. In this table, the case of exactly 85 degrees of freedom is not given, but by interpolating between the case of 90 and that of 80 degrees of freedom, we can find that it would take a coefficient equal to .211 to be significant and one of .275 to be very significant. Our obtained r of .61 is decidedly above the latter. In the first problem mentioned in this chapter, where $N = 10$, there are 8 degrees of freedom. Here it would take an r of .632 to be significant and one of .765 to be very significant. Our obtained r of .76 just misses the "very significant" category, the r of $-.69$ is significant but not very significant, and the r of .14 is decidedly insignificant. It can be seen that the second test is very direct and easy to apply, and wherever a test of the null hypothesis is to be made, it is much to be preferred.

REGRESSION EQUATIONS

The Meaning of a Regression Equation.—The main use of a regression equation is to predict the most likely measurement in one variable from the known measurement in another. If the correlation between Y and X were perfect (with a coefficient of $+1.00$ or -1.00), we could make predictions of Y from X or of X from Y with maximum accuracy; the errors of prediction would be zero. If the correlation were zero, no predictions are possible. Between these two limits, predictions are possible with varying degrees of accuracy. The higher the correlation the greater the accuracy of prediction or the smaller the errors of prediction.

In the preceding chapter (pages 193*f.*), we predicted that the most likely measurement in Y corresponding to a measurement in X was the mean of the cases in the column. We were thus able to predict only for the X values at the midpoints of the class intervals in X . We should also like to be able to predict not only for midpoints but also for all values of X . This the regression equation enables us to do. We found (see Figs. 31 and 32) that the means of the columns (and of the rows) tended to lie along a straight line, with some minor deviations from strict linearity. We shall now assume that the best predictions of Y from X do lie along a line that best fits the means of the columns when those means are weighted according to the number of cases represented in each one. This is known as the *line of best fit*, or the *regression line*. When predicting Y from X , we have one such line, for the regression of

Y on X ; and when predicting X from Y , we have another such line, for the regression of X on Y . The two regression lines for the achievement-test data will be found in Fig. 37. Only when a correlation is perfect will the two lines coincide throughout their lengths. The higher the correlation, plus or minus, the closer together they tend to lie. All such pairs of regression lines intersect at the point representing the means of Y and X ; in this case, they cross at $X = 78.15$ and $Y = 115.28$.

The Regression Equations and Regression Coefficients.—From elementary algebra, the student should remember that the equation for a straight line, in general form, is $Y = a + bX$. Such an equation

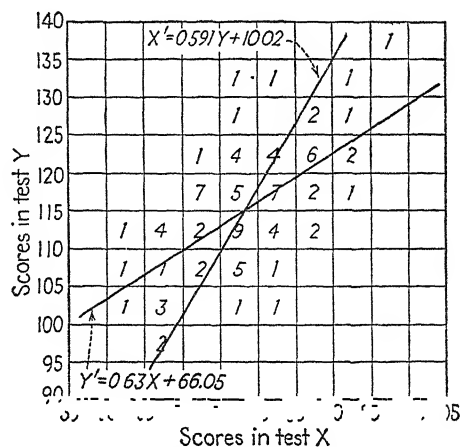


FIG. 37.—A scatter diagram for two examinations, with the two regression lines represented.

completely describes a line when a and b are known; they are the regression coefficients and must be obtained from the data we have. Leaving out of account for the moment the coefficient a , we should have $Y = bX$, or Y equals b times X . We see from this that b is a ratio, and it tells us how many units Y is increasing for every increase of one unit in X . If b were 2, then for every unit of increase in X , Y increases two units. If $b = 0.5$, then for every unit increase in X , Y increases a half unit. The b coefficient gives us the *slope* of the regression line, and it depends upon the coefficient of correlation and the two standard deviations, as in the formula

$$b_{yx} = r_{yx} \left(\frac{\sigma_y}{\sigma_x} \right) \quad (61)$$

where b_{yx} , with the subscripts in that order, implies that we are predicting Y from X and where this is also true for r_{yx} .

When we want to predict X from Y , we have a different regression equation with a different b , which is given by the formula

$$b_{xy} = r_{xy} \left(\frac{\sigma_x}{\sigma_y} \right) \quad (62)$$

The coefficient of correlation is, of course, numerically the same in both cases, since $r_{yx} = r_{xy}$. But in each case, the b 's are different and are equal to r times the ratio of the standard deviation of the *predicted* variable to that of the variable *predicted from*. We frequently speak of the predicted variable as the *dependent* variable and of the one predicted from as the *independent variable*. The reason for this is that in predicting Y from X , we arbitrarily take any value of X that we wish at the moment, whereas the Y we predict from it *is* dependent upon what X we have chosen. Once we have picked out a certain X , Y is immediately fixed by our regression equation.

The regression coefficient a is merely a constant that we must always add in order to take account of the fact that our two means are not the same value. It is given by the formulas

$$a_{yx} = M_y - (M_x)b_{yx} \quad (63a)$$

$$a_{xy} = M_x - (M_y)b_{xy} \quad (63b)$$

where the first one concerns the equation for the regression of Y on X and the second concerns the equation for the regression of X on Y .

The derivation of the entire regression equation is more often accomplished by one composite formula, combining the derivations of a and b into one operation as follows:

$$Y' = r \left(\frac{\sigma_y}{\sigma_x} \right) (X - M_x) + M_y \quad (64a)$$

and

$$X' = r \left(\frac{\sigma_x}{\sigma_y} \right) (Y - M_y) + M_x \quad (64b)$$

We use Y' and X' here rather than Y and X to show that they are predicted rather than obtained values.

Applying these formulas to the illustrative data on achievement examinations, we have

$$\begin{aligned}
 Y' &= .61 \left(\frac{7.85}{7.60} \right) (X - 78.15) + 115.28 \\
 &= (.61)(1.03)(X - 78.15) + 115.28 \\
 &= .630X - 49.23 + 115.28 \\
 &= .630X + 66.05 \\
 X' &= .61 \left(\frac{7.60}{7.85} \right) (Y - 115.28) + 78.15 \\
 &= .591Y + 10.02
 \end{aligned}$$

Interpreting these equations, we may say that Y' increases .63 units for every unit increase in X and that X' increases .591 units for every unit increase in Y . One way of checking the accuracy of the regression equations as a whole is to substitute M_x in the first one to see whether Y' is the mean of the Y 's and to substitute M_y in the second to see whether we obtain M_x as our prediction of X . Another check as to the accuracy of computation of the b coefficients is the equation

$$b_{yx}b_{xy} = r^2 \quad (65)$$

In other words, the product of the two b coefficients is equal to the square of the coefficient of correlation. In this instance

$$(.63)(.591) = .3723 = .61^2$$

Regression Coefficients from Ungrouped Data.—When data have not been grouped in class intervals, the derivation of the b coefficient requires another formula, which reads

$$b_{yx} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma X^2 - (\Sigma X)^2} \quad (66)$$

When this formula is applied to the data in Table 62 (page 204), we have

$$\begin{aligned}
 b_{yx} &= \frac{4,720 - 4,550}{6,240 - 4,900} \\
 &= \frac{170}{1,340} \\
 &= .127
 \end{aligned}$$

The a coefficient is obtained by means of formula (63a) and is solved as follows:

$$\begin{aligned}
 a_{yx} &= 6.5 - (7.0)(.127) \\
 &= 6.5 - .89 \\
 &= 5.61
 \end{aligned}$$

The regression equation is therefore $Y' = 5.61 + .127X$. The equation for the regression of X on Y can be obtained by similar operations, substituting Y for X , and vice versa, in formula (66). The solution is

$$\begin{aligned} b_{xy} &= \frac{4,720 - 4,550}{5,330 - 4,225} \\ &= \frac{170}{1,105} \\ &= .154 \end{aligned}$$

And

$$\begin{aligned} a_{xy} &= 7.0 - (6.5)(.154) \\ &= 7.0 - 1.0 \\ &= 6.0 \end{aligned}$$

Checking the b coefficients, $b_{yx}b_{xy} = (.127)(.154) = .0196 = r^2$, which is in agreement with r^2 as previously known (see page 204).

Predictions from the Regression Equations.—As an illustration of how a regression equation is applied in prediction, let us assume some values of X and find the corresponding Y' values. Because in the preceding chapter we predicted Y 's corresponding to midpoints of the intervals of X , let us do the same here for the sake of comparison, remembering that we might have chosen any values of X that we pleased. Table 65 gives the X values and their corresponding Y'

TABLE 65.—PREDICTIONS OF Y FROM X AND X FROM Y BY MEANS OF REGRESSION EQUATIONS*

$$Y' = 0.63X + 65.95$$

If $X =$	62	67	72	77	82	87	92	97
$Y' =$	105.1	108.3	111.4	114.6	117.7	120.9	124.0	127.2
$M_c =$	107.0	105.5	114.9	114.5	116.4	120.3	124.0	132.0

$$X' = 0.59Y + 10.29$$

If $Y =$	97	102	107	112	117	122	127	132	137
$X' =$	67.3	70.3	73.3	76.2	79.2	82.1	85.1	88.0	91.0
$M_{row} =$	67.0	70.3	74.0	75.9	78.6	83.2	85.8	83.7	97.0

* The data involved are from the two examinations correlated in Table 64. The means of the columns and rows are obtained from Table 59, p. 193.

values. When X is 62, Y' is 105.1, and when $X = 97$, $Y' = 127.2$, etc. It is interesting to compare these particular predictions with the means of the columns, which are given in the third row of Table 65. The discrepancies will be found very small as a rule. Granting that the column means are generally not very reliable because of small

samples, we may feel more assurance in the Y' predictions because they are determined from the trend of the entire data rather than by small samples in separate columns. The predictions of X' from Y are given in the second section of Table 65 and are compared with the means of the rows as a matter of interest.

As a practical means of prediction, a graphic method will often be the most suitable procedure. If the regression lines are drawn as in Fig. 37 on cross-section paper, for any value of X on the base line, one can follow vertically up to the regression line and note the corresponding Y value at this point. One can read to the nearest unit with sufficient accuracy for practical work. The drawing of the regression line is simple in that two points determine the position of a line. One point can be at the two means, which will serve for both regressions. Another point for the regression of Y on X might be at $X = 60$, $Y = 103.85$; and a third point, for checking purposes, might be at $X = 100$, $Y = 129.05$. For the regression of X on Y , points might be located conveniently at $Y = 100$, $X = 69.12$, and $Y = 130$, $X = 86.85$.

Standard Errors of the Estimates.—In the preceding chapter, we saw that the errors of prediction ($Y - Y'$ in the one case and $X - X'$ in the other) can be squared, summed, averaged and then the square root extracted in order to obtain the standard error of the discrepancies between observed values and predicted values. There we derived the standard error of the estimate from the discrepancies themselves; here we shall now see that it is not necessary to compute the errors of prediction. When we have predicted on the basis of regression equations, we can estimate the margin of error of prediction, as given by σ_{yx} (or by σ_{xy}) from the coefficient of correlation. The formulas are

$$\sigma_{yx} = \sigma_y \sqrt{1 - r^2_{yx}} \quad (67a)$$

and

$$\sigma_{xy} = \sigma_x \sqrt{1 - r^2_{xy}} \quad (67b)$$

in both of which the terms are now well known. It will be seen that the two equations are the same except for the use of σ_y when we are predicting Y and of σ_x when we are predicting X (for $r_{yx} = r_{xy}$). The two standard deviations are multiplied by the common factor $\sqrt{1 - r^2}$. This factor is always less than 1.00 and gives us an estimate of the reduction in errors of prediction from knowledge of correlated measurements as compared to errors of prediction without that knowledge. When r is zero, this element equals 1.00, and then $\sigma_{yx} = \sigma_y$, and $\sigma_{xy} = \sigma_x$. In other words, when $r = 0$, there is no basis for prediction. When $r = 1.0$ (or -1.0) the element reduces to zero, and so does the standard

error of estimate. This coincides with the expectation that the margin of error of prediction is zero when the correlation is perfect.

The interpretation of the SE of the estimate when r is neither zero nor 1.00 is somewhat as follows. Like any SE , σ_{yx} can be referred to the normal curve of distribution. For the examination problem,

$$\sigma_{yx} = 7.85 \sqrt{1 - .3721} = 6.22$$

and

$$\sigma_{xy} = 7.60 \sqrt{1 - .3721} = 6.02$$

No matter in what part of the measuring scale we are predicting (within the range of obtained scores, naturally) we assume that the

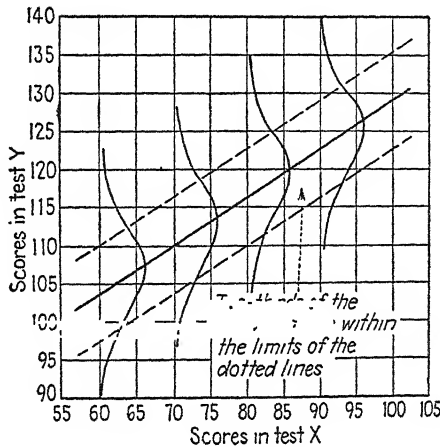


FIG. 38.—The line of regression of Y upon X , showing the range of observed values expected in Y at any value of X . Parallel dotted lines are drawn above and below regression line at a distance of one standard error of the estimate each way.

margin of error is the same. When we predict Y from X , the average dispersion of observed measurements about Y' is given by a σ of 6.22. We expect two-thirds of the observed cases to lie within the limits of plus or minus 6.22 from Y' . This situation is illustrated graphically in Fig. 38. There we have the regression line, along which the predicted Y 's lie, and in dotted lines we have the limits of one SE of the estimate on either side of it. Had we plotted a point for every individual, we should have expected about two-thirds of them to fall between the two dotted lines. To make a particular prediction, when $X = 90$, $Y = 122.8$. The odds are 2 to 1 that any individual whose X -score is 90 will not fall below 116.6 or go above 129.0. We could state other

odds for a divergence of 2σ either way or any other distance. It all depends upon our purposes.

We can prepare a similar diagram showing the limits of the middle two-thirds of the individuals about the regression of X on Y , and we can interpret the errors of prediction in a similar manner. It will be noted that the margin of error as given by σ_{xy} is 6.02, or 0.2 smaller in predicting in the other direction, *i e.*, X from Y , but this is merely because σ_x is smaller than σ_y . The *percentage* of error is the same in the two cases. The ratio of σ_{yx} to σ_y is exactly the same as the ratio of σ_{xy} to σ_x , and that ratio is given by the factor $\sqrt{1 - r^2}$. This factor we meet again with a name attached to it (page 222).

The Reliability of a Regression Coefficient.—The b coefficient in the regression equation has its sampling error like all statistics. This is given by

$$\sigma_{b_{yx}} = \frac{\sigma_{yx}}{\sigma_x \sqrt{N}} \quad (68)$$

or by

$$\sigma_{b_{yx}} = \frac{\sigma_y}{\sigma_x} \sqrt{\frac{1 - r^2}{N}} \quad (69)$$

The SE for b_{xy} would be the same, except for changing the x and y subscripts around. For our examination problem

$$\begin{aligned} \sigma_{b_{yx}} &= \frac{6.22}{(7.60)(9.3274)} \\ &= \frac{6.22}{70.88824} \\ &= .088 \end{aligned}$$

We may say that the odds are 2 to 1 that the obtained b_{yx} of .63 does not deviate from the true b_{yx} by more than .088; consequently the true coefficient is probably (odds 2 to 1) not less than .542 or higher than .718. There is practically no chance that the true b coefficient here is zero or less.

INTERPRETATIONS OF A COEFFICIENT OF CORRELATION

How High Is Any Given Coefficient of Correlation?—We have said or implied that any coefficient of correlation that is significantly not zero denotes some degree of relationship between two variables. But we need further orientation on the matter, for the strength of relationship can be regarded from a number of points of view, and it is not correct from any one of these points of view to say that the degree of

relationship is exactly proportional to the size of r . The coefficient of correlation does *not* give directly anything like a percentage of relationship. We cannot say that an r of .50 indicates two times the relationship of that which is indicated by an r of .25. Nor can we say that an increase in correlation from $r = .40$ to $r = .60$ is equivalent to an increase from .70 to .90. The coefficient of correlation is an index number, not a measurement on a linear scale.

A General Verbal Description of Coefficients.—Our interpretation of the size of r depends very much upon what we propose to do with it or the reasons why we computed it. What would be a large correlation coefficient for one purpose would be regarded as a small one for another. Interpretation is therefore largely a relative matter; relative to the area of investigation in which we are working and to other factors. But taking correlations just at large, without particular regard to their use and as a general orientation, we may say that the strength of relationship can be described roughly as follows for various r 's:

Less than .20	Slight; almost negligible relationship
.20–.40	Low correlation; definite but small relationship
.40–.70	Moderate correlation; substantial relationship
.70–.90	High correlation; marked relationship
.90–1.00	Very high correlation; very dependable relationship

It should be said that the coefficients given should be interpreted as stated only when, by comparison with the standard error of r , they prove to be significant. It should also be said that the same interpretations apply alike to negative and positive r 's of the same numerical size.

Particular Uses Have a Bearing on Interpretation of r .—The general descriptive list should be qualified by reference to particular uses of r . A common use is to indicate the agreement between aptitude tests and scholastic or vocational success. This refers to a practical *validity coefficient*, of which more will be said in a later chapter. Common experience shows that validity coefficients for mental tests range from about .30 to .80. Those who employ tests in guidance and selection feel that a correlation should be at least .45 for material usefulness and that the best results come when r is above .60. There also seems to be some agreement that the *reliability* of a test is not sufficiently high for individual predictions unless the self-correlation of the test is above .90, preferably .95, and even for group predictions or for research on groups, a reliability coefficient above .80 is usually demanded. These standards are not always attainable, however, and one may well use tests of lower validity and reliability, but with increased caution

consistent with the greater margin of error of prediction involved. The experienced counselor can and does use tests in guiding individuals when the reliability of those tests is below .90. When validity is high (above +.60), the counselor need not be very much concerned about reliability.

When one is investigating a purely theoretical problem, even very small correlations, if statistically significant (undoubtedly not zero), are often very indicative of a psychological law. Whenever a relationship between two variables is established beyond reasonable doubt, the fact that the correlation coefficient is small may merely mean that the measurement situation is contaminated by many things uncontrolled or not held constant. One can readily conceive of an experimental situation in which, if all irrelevant factors had been held constant, the r might have been 1.00 rather than .20. For example, the correlation between an ability score and scholarship is .50, since both are measured in a population whose scholarship is also allowed to be determined by effort, attitudes, marking peculiarities of the instructors, and what not. Were all the other determiners of scholarship held constant and were both aptitude and marks perfectly measured, the r would be 1.00 rather than .50. This line of reasoning indicates that where any correlation between two things is established at all, and particularly where there is a causal relationship involved, the fundamental law implies a perfect relationship. Thus, in nature, correlations of zero or 1.00 are the rule between variables when isolated. The fact that we obtain anything else is because of the inextricable interplay of variables that we cannot measure in isolation.

The practical conclusion from this is that *a correlation is always relative to the situation under which it is obtained, and its size does not represent any absolute natural or cosmic fact.* To speak of the correlation between intelligence and scholarship is absurd. One needs to say *which* intelligence, measured under *what* circumstances, in *what* population, and to say *what kind* of scholarship, measured by *what* instruments, or judged by *what* standards. *Always, the coefficient of correlation is purely relative to the circumstances under which it was obtained and should be interpreted in the light of those circumstances; never, certainly, in any absolute sense.*

How much faith one should place in any relationship shown by a coefficient of correlation also depends upon the urgency of the outcome. There are probably many medical treatments, such as some inoculations, vaccines, and the like, concerning which the knowledge is rather incomplete, which are administered even though the correlation between

the treatment and living (or between non-treatment and dying) is of the order of .10 to .20. Although the probabilities of living may be increased by only 1 per cent by the treatment, the saving of 1 life in 100 is regarded as worth the effort. If a procedure in education promised only 1 per cent improvement over guesswork, we should pay little attention to it, because the seriousness of the outcome would not justify the means. It may be said in passing, however, that failures to predict in vocational and educational practice are more generally recognized by reason of correlational checkup than are failures to predict in medical practice, where correlational checkup is less often

TABLE 66—INDICATORS OF THE IMPORTANCE OF COEFFICIENTS OF CORRELATION

r_{xy}	k_{xy} Coefficient of alienation	$100(1 - k_{xy})$ Percentage reduc- tion in errors of prediction of Y from X	$100r^2_{xy}$ Percentage of variance accounted for
00	1 000	0.0	0 00
05	999	1	0 00
10	.995	.5	1 00
15	989	1 1	2.25
.20	980	2 0	4 00
25	968	3 2	6 25
30	954	4 6	9 00
35	937	6.3	12 25
.40	917	8 3	16 00
.45	893	10.7	20 25
.50	866	13 4	25 00
.55	.835	16 5	30 25
60	800	20 0	36.00
.65	760	24 0	42 25
.70	.714	28 6	49 00
.75	661	33 9	56.25
80	600	40.0	64.00
.85	.527	47.3	72 25
.90	.436	56 4	81 00
.95	.312	68 8	90 25
98	.199	80 1	96 00
.99	.141	85.9	98 00
995	100	90 0	99 00
.999	.045	95 5	99 80

made. In addition to the difference in relative seriousness of the outcomes of prescription in the two cases, this factor of better knowledge of goodness of results may be an important reason for the higher standards of prescriptive accuracy demanded in education than are sometimes required in other fields.

The Coefficient of Alienation.—Whereas r indicates the strength of relationship, we also have the *coefficient of alienation*, k , to indicate the degree of *lack* of relationship. By formula

$$k = \sqrt{1 - r^2} \quad (70)$$

Squaring both sides of this equation, we have

$$k^2 = 1 - r^2$$

And transposing, we have

$$k^2 + r^2 = 1.00$$

Thus, although we might have expected k plus r to equal 1.00, it is rather the sum of their squares that equals 1.00. If r is .50, k is *not* also .50 but .866. When r is .50, then, the degree of relationship is less than the degree of *lack* of relationship. It is only when $r = .7071$ that we have a balance between relationship and lack of relationship, for k also then equals .7071. Then $r^2 + k^2 = .50 + .50 = 1.00$. Other values of k for different sizes of r can be found in Table 66. Sometimes we wish to stress the point of independence between two things rather than their closeness of agreement. In such instances, we present k as well as r .

The Index of Forecasting Efficiency.—In the formula for the *SE* of the estimate, $\sigma_{yx} = \sigma_y \sqrt{1 - r^2_{yx}}$, we can now see that the factor under the radical, $\sqrt{1 - r^2_{yx}}$, is really the coefficient of alienation. We could rewrite the formula as $\sigma_{yx} = \sigma_y k_{yx}$. If we were to multiply k by 100, we should have the percentage σ_{yx} is of σ_y . When $r = .61$, as in our recent illustration, $k = .7924$. The *SE* of the estimate in this problem is 79.24 per cent of the observed dispersion of observations. Our margin of error in predicting Y *with* knowledge of X scores is about 79 per cent as great as the margin of errors we would make *without* knowledge of X scores. For then we predict every Y to be the mean of the Y 's, and the *SE* of the prediction then equals σ_y . The *reduction* of our margin of error is 100 minus 79.23, or 20.77 per cent. The *index of forecasting efficiency* is defined as the percentage reduction in errors of prediction by reason of correlation between two variables. The general, simplified formula is

$$E = 100(1 - \sqrt{1 - r^2}) \quad (71)$$

or

$$E = 100(1 - k)$$

The calculation of E is facilitated by Table 66, where many of the E values are given for corresponding r 's. Inspection will show that r must be as high as about .45 before E is 10 per cent, an efficiency that is regarded as about the lower limit of usefulness for mental tests. The better tests, with validity coefficients of .60, have an E of 20 per cent, and the best tests, when $r = .75$, have an E of about 34 per cent. Although these efficiencies seem small, we must treat them in a relative, not an absolute sense. It is probable that the efficiency of predictions based upon the average unsystematic interview is around 5 per cent. With this as our base, the picture of efficiency of tests looks much better.

The Coefficient of Determination.—A final mode of interpretation of r is in terms of r^2 , which is called the *coefficient of determination*. This coefficient, or r^2 , gives us (when multiplied by 100) the percentage of the *variance* (see page 146) in Y that is associated with or determined by variance in X . When $r = .50$, the percentage of the variance in Y that is accounted for by variance in X is 25, or one-fourth. To account for half the variance of any set of measurements, the r with another variable would have to be .7071. The percentage of the variance in Y *not* determined by or associated with variance in X is given by $100k^2$, which is called the *coefficient of non-determination*. These statements about determination of Y by X are reversible and apply equally well to determination of X by Y . We should speak of "determination" of one thing by another, however, only when a causal relationship can be logically defended; otherwise the expression "associated with" or "accounted for" (by way of prediction) is better. In Table 66, several of the $100r^2$ values are given for corresponding r 's.

ASSUMPTIONS UNDERLYING THE PRODUCT-MOMENT CORRELATION

The student should be warned before leaving this chapter concerning the restrictions that should be observed in the use of the Pearsonian coefficient of correlation. The most important requirement for the legitimate use of the Pearson r is that the trend of relationship between Y and X be rectilinear, in other words, a straight-line regression. This can be determined, as a rule, by inspection of the scatter diagram. If the distribution of the cases within the correlation diagram appears to be elliptical, without any indications of a decided bending of the ellipse, the chances are that the relationship is rectilinear. Even if it is

not, the deviation from a straight-line relationship may be so slight that we may assume rectilinearity as a first approximation, and the degree of correlation indicated by r will be fairly close to any index of correlation like the *correlation ratio* (see page 231) that is applied when there is curvature in the trend. When there is an obvious bending of the distribution of cases, a correlation ratio is indicated as the best index of correlation.

But there are in educational and psychological measurements certain factors that produce artificially curved scatters in the correlation diagram. This may happen when one or both distributions taken alone are badly skewed and the skewing is produced artificially by the faulty measuring scale, with its systematically shifting unit of measurement. If there is good reason to believe that this may be the case, it is best to normalize the skewed distribution by means described in Ch. VII. When distributions are corrected for skewness, the curvature in the regression is frequently eliminated, and linearity then obtains. If curvature still remains, then the Pearson r is not to be used to indicate the amount of correlation.

There is nothing in what has been said to demand that the Pearson r is to be computed only with normal distributions. The forms of distribution may be various, even rectangular. The important consideration is whether or not in all columns the dispersions are approximately equal, as indicated by the column standard deviations and also in all rows. This condition goes by the name *homoscedasticity*. When columns (and rows) are relatively homoscedastic, we may compute a Pearson r and its standard error. This condition will prevail generally when the two distributions are fairly symmetrical within themselves; so we need not go so far as to compute standard deviations of columns and rows in order to find out. It is when distributions are markedly skewed that significant departures from homoscedasticity occur.

Exercises

- 1 Using the first 10 cases in Data AA, compute a Pearson r between ascendance-submission scores and masculinity-femininity scores, using formula (54). Find a similar correlation between the same two traits, using the last 10 cases. Compare the two r 's, and draw conclusions.
- 2 Correlate the first 10 sets of scores in inferiority feelings and nervousness (Data AA), using formula (56). Do the same for the last 10 cases in the same two traits. State conclusions.
- 3 Prepare a scatter diagram for the correlation of ascendance-submission scores with masculinity-femininity scores. Compute the product-moment r , using formula (55). Determine the reliability of the coefficient.

DATA 44 — SCORES IN FOUR TRAITS, DERIVED FROM PERSONALITY QUESTIONNAIRES

Ascendance- submission	Masculinity- femininity	Inferiority feelings	Nervousness
23	52	2	5
7	40	9	12
16	47	9	10
21	49	6	15
23	48	5	11
11	44	10	4
16	39	9	13
19	41	3	0
24	42	2	6
13	42	6	12
18	30	10	16
20	53	6	9
19	40	9	17
23	42	5	16
30	58	9	13
38	46	7	11
23	34	5	13
16	42	7	15
8	56	13	12
21	51	11	7
20	48	3	7
10	43	12	14
20	44	3	2
10	48	10	16
12	51	7	7
17	41	9	11
18	34	7	6
14	43	10	11
17	62	6	16
30	49	4	15
15	40	9	10
22	42	9	12
17	46	7	12
10	47	10	16
23	47	3	8
17	42	7	17
18	52	4	4
9	45	9	9
21	35	7	6
10	39	4	8
21	44	4	6
22	55	3	7
24	52	3	8
17	50	5	9
29	62	4	7

CHAPTER XII

OTHER CORRELATION METHODS

Pearson's product-moment coefficient is the standard index of the amount of correlation between two things, and we employ it whenever it is possible and convenient to do so. But there are data to which this kind of correlation method cannot be applied, and there are instances in which it can be applied but in which, for practical purposes, other procedures are more expedient. The Pearson coefficient cannot or should not be computed, for example, unless the two variables X and Y are measured on continuous metric scales and unless the regressions are linear. Many of our data are in terms of frequencies of cases having certain attributes; they are variables of a "qualitative" rather than a quantitative sort. Less often, two continuously measured variables bear to one another a relationship that is curved rather than in the form of a straight line. In this chapter will be described some procedures that take care of these irregular situations and of other situations where short-cut methods are better used to compute a Pearson r or its equivalent.

SPEARMAN'S RANK-DIFFERENCE CORRELATION METHOD

Probably the most commonly known and commonly employed procedure applied to regular data in the place of the product-moment method, is the rank-difference method of Spearman. It is conveniently applied as a quicker substitute when the number of pairs, or N , is less than 50. It is even more conveniently applied when the data are already in terms of rank orders rather than in terms of measurements.

The Computation of a Spearman Rho.—If we have data in terms of measurements or scores, it is first necessary to translate them into rank orders. The procedure will be demonstrated by means of the data in Table 67. There we have 15 pairs of scores for 15 individuals who responded to sets of cartoons and limericks by judging their humor values, each on a 5-point scale. The score in each case is the sum of the points each individual assigned to the set. We could correlate these scores in the usual manner, described in the preceding chapter.

TABLE 67.—A RANK-DIFFERENCE CORRELATION BETWEEN HUMOR SCORES IN REACTIONS TO CARTOONS AND TO LIMERICKS

Cartoon score	Limerick score	R_1	R_2	D	D^2
47	75	11	8	3	9 00
71	79	4	6	2	4 00
52	85	9	5	4	16 00
48	50	10	14	4	16 00
35	49	14 5	15	0 5	0 25
35	59	14 5	15	0.5	0 25
41	75	12 5	8	4.5	20 25
82	91	1	3	2	4 00
72	102	3	1	2	4.00
56	87	7	4	3	9.00
59	70	6	10	4	16 00
73	92	2	2	0	0 00
60	54	5	13	8	64 00
55	75	8	8	0	0 00
41	68	12 5	11	1 5	2 25
					165 00
					ΣD^2

The rank-difference method will be found shorter. The following steps are necessary:

Step 1. Rank the individuals from the highest to the lowest in the first variable (here it is "cartoon score"), and call these ranks R_1 . The highest score receives the rank of 1 (which is arbitrary; we might have called it 15), the next highest 2, etc. The only difficulty encountered is when we find ties. For example, in Table 67, two individuals have scores of 41. One of them comes at rank 12 and the other at rank 13. We do not know which, if either, is better, yet we must fill these two rank positions; so we take the average of the tied ranks and call them both 12.5. We make certain that the next ranking scorer is called 14, unless he also is tied. We find that he is tied with another who has a score of 35. We treat these two in a similar manner; so they become each 14.5. If the lowest person is not tied with others, the last rank should be equal to N (in this case, 15). This serves as a check as to accuracy of ranking, though, of course, it will not detect inversions in rank order somewhere along the line.

It merely shows whether any rank has been repeated, whether any individuals have been overlooked, or whether ties have somewhere not been properly treated.

- Step 2. Rank the second list of measurements in a similar manner, and call them R_2 . In this problem, there are three scores of 75 for the individuals occupying places 7, 8, and 9. We call them all 8, leaving out of the list 7 and 9. This treats the three alike, as they should be, yet gives us a full set of 15 ranks.
- Step 3. For every pair of ranks (for each individual), determine the difference in ranks. The smaller one can be subtracted from the larger one in each case, with no attention being paid to algebraic signs, for they are all going to be squared anyway.
- Step 4. Square each difference to find D^2 .
- Step 5. Sum the squares of the differences (see the last column of Table 79) to find ΣD^2 . The sum in our illustrative problem is 165.00
- Step 6. Compute the coefficient ρ (Greek letter rho) by means of the formula

$$\rho = 1 - \frac{6\Sigma D^2}{N(N^2 - 1)} \quad (71)$$

where ΣD^2 = sum of the squared differences between ranks.

N = number of pairs of measurements.

In this problem

$$\begin{aligned} \rho &= 1 - \frac{6 \times 165}{15 \times 224} \\ &= 1 - .295 \\ &= .705, \text{ or } .70. \end{aligned}$$

By this procedure, then, the estimate of the amount of correlation between the two sets of scores is .70. How shall we interpret this correlation, as compared with a Pearson r ?

TABLE 68.—TABLE FOR CONVERTING A SPEARMAN RHO COEFFICIENT INTO ITS EQUIVALENT PEARSON r

ρ	r	ρ	r	ρ	r	ρ	r
.05	.052	.30	.315	.55	.568	.80	.813
.10	.105	.35	.364	.60	.618	.85	.861
.15	.157	.40	.416	.65	.668	.90	.908
.20	.209	.45	.467	.70	.717	.95	.954
.25	.261	.50	.518	.75	.765	1.00	1.000

Interpretation of a Rho Coefficient.—The rank-difference coefficient is practically equivalent to the Pearson r numerically. There is a conversion formula by which the corresponding Pearson r can be estimated from rho. But this formula assumes large samples, which is precisely what we do not have when we compute rho, and in no case is the difference between rho and r greater than .018, and in every case, except for coefficients of zero or 1.00, r is greater than rho. We may therefore treat an obtained rho as an approximation to r and under these circumstances interpret the outcome of a correlation study accordingly.¹

The reliability of rho, as indicated by its standard error, is in much doubt, but a rough approximation is given by the formula

$$\sigma_{\rho} = \frac{1.05(1 - \rho^2)}{\sqrt{N - 1}} \quad (72)$$

For the illustrative problem

$$\begin{aligned} \sigma_{\rho} &= \frac{1.05(1 - .497025)}{\sqrt{14}} \\ &= .14 \end{aligned}$$

The obtained correlation is highly significant in that it is so far removed from zero. The fluctuation of other sample rho coefficients about some true value of rho for this type of data would be rather narrowly delimited. Allowing a t ratio of 3.0 and accepting .14 as our standard error, we can surmise that the true correlation lies somewhere between .28 and 1.00.

A Method of Dealing with Ties.—DuBois has shown that the procedure of giving tied scores a common rank equal to the mean of the ranks involved in the ties is a good approximation, but that when more than two or three scores are tied, a better estimate is desirable. His formula is

$$R_c = \sqrt{M^2_R + \frac{n^2 - 1}{12}} \quad (73)$$

where R_c = corrected rank for the ties.

M_R = mean of the ranks for the ties.

n = number tied.

Every case of ties would have to be treated separately. For example, in our second set of scores in Table 67, there are three ties for eighth

¹ Table 68 gives a quick means of converting rho into the corresponding Pearson r .

place (or three identical scores for places 7, 8, and 9). Applying DuBois's formula¹

$$\begin{aligned} R_c &= \sqrt{8^2 + \frac{9-1}{12}} \\ &= \sqrt{64 + .67} \\ &= \sqrt{64.67} \\ &= 8.04 \end{aligned}$$

Had we used 8.04 instead of 8 in the computation of rho, the result would hardly have been affected. With five or more ranks tied, the correction would probably have made some material difference.

THE CORRELATION RATIO

The correlation ratio is a very general index of correlation particularly adapted to data in which a curved regression prevails. Among test scores, linear relationships are apparently the almost universal type of regression. Normality, or near normality, in both distributions correlated is almost sufficient in itself to promote linearity. Outside the sphere of psychological and educational tests, however, or when outside variables are correlated with test scores, we sometimes encounter curved trends in the scatter diagram. The means of the columns do not progressively increase as we go up the *X*-scale. They may increase slowly at first then rapidly later, or they may increase to a maximum in the center and then decrease, or other systematic divergencies from linearity may be apparent.

Non-linear Regressions.—A common instance of non-linear relationship is found when we correlate performance scores with chronological age. Typically, goodness of performance as measured, increases most rapidly from ages five to ten and thereafter shows a slackening in upward trend through the teens. If we follow the progression still further, we find typically a maximal performance somewhere in the twenties, with slow decline to the forties and an increasing rate of decline thereafter. If we included all ages from five to seventy-five in our correlation study and if we computed the usual Pearson *r* between age and scores, the *r* would probably prove to be near zero. On such a correlation diagram, the scattering of points would be considerably dispersed from any straight line that we might try to draw through the data, slanting upward or slanting downward. Inspection would show, nevertheless, a law of relationship between age

¹ DuBois, P. H., Formulas and tables for rank correlation. *Psychol. Rec.*, 1939, 3, 46-56.

and performance but a relationship that takes into account the waxing and waning of ability both within the span of ages studied

We might break the chart in two and treat by themselves the years during which there is improvement and by themselves the years during which there is decline. We should be able to compute a positive correlation for the earlier span and a negative correlation for the later span by assuming straight-line trends. But these would be of doubtful significance and certainly would not do justice to the full strength of relationship, even within the two segments of life span. The reason is that the trends still deviate from straight lines. Curvature has been

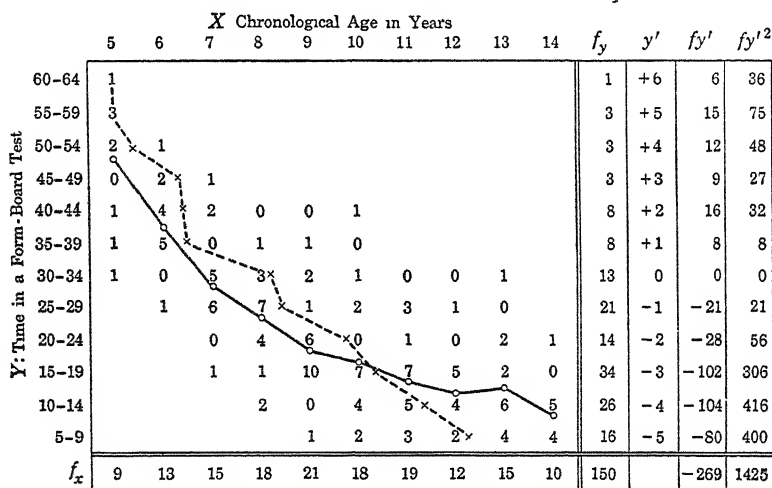


FIG. 39—A scatter diagram for a correlation-ratio problem

overlooked, and to that extent the index of correlation is perhaps markedly underestimated.

Two Regression Lines and Two Correlation Ratios.—The scatter diagram in Fig. 39 represents a sample of relationship between performance score in a form-board test and chronological age between five and fifteen years inclusive. Here the score is time required for completion, and so a high number indicates a poor performance, and the trend is downward. But the relationship obviously drops most rapidly during the first 3 years and settles down to slight changes from year to year during the last 3 years. Two regression lines are drawn in the diagram to show more clearly the trends. The regression of test score on age is shown by the solid line that is drawn connecting the circlets, which are plotted at the means of the columns. The regression

for age upon test score is shown by the dotted line and the means of the rows, by the x 's. Just as we found two regression lines (for an imperfect correlation) in the last chapter, where linear regressions were involved, so here we find two regression curves, differing in shape as well as in slope. We have accordingly two correlation ratios, or eta coefficients, one for each of the regressions, and they will not necessarily be the same in value. This result differs from that in the case of linear correlation, where $r_{yx} = r_{xy}$. The two correlation ratios are given by the formulas

$$\eta_{yx} = \frac{\sigma_{y'}}{\sigma_y} \quad (74a)$$

and

$$\eta_{xy} = \frac{\sigma_{x'}}{\sigma_x} \quad (74b)$$

where η_{yx} = correlation ratio for the regression of Y on X , and η_{xy} = its analogue.

$\sigma_{y'}$ = standard deviation of the values (Y') predicted from X ,
and $\sigma_{x'}$ = standard deviation of the X values predicted from Y .

σ_y and σ_x = standard deviations of the two total distributions.

The manner in which $\sigma_{y'}$ and $\sigma_{x'}$ are determined will next be explained.

The Computation of a Correlation Ratio.—Remember that in a prediction problem of this sort, the best prediction of Y for any column is the mean of the Y 's in that column. This prediction will have the smallest sum of squared deviations from the observed Y 's in that column. So Y' for each column is the mean of that column. We therefore first compute the means of the columns. These are listed in column (3) in Table 69. Now if there were no correlation, no law of relationship between Y and X , these Y' values would lie along the level of the mean of all the Y values, which in this problem is 23.0. No predictions could then be made on the basis of knowledge of X values. For every column with its X value (midpoint), the most probable corresponding Y would be 23.0, and our margin of error would be indicated by σ_y . It would be as large as if we had no knowledge of X for each individual (see Ch. X for a more complete discussion of this point).

The more the means of the columns deviate from the mean of all the Y 's the more accurate our predictions are. We are therefore interested in how far the Y' values do deviate from 23.0 in this problem. Those discrepancies ($Y' - M_y$) are given in column (4) of Table 69. As usual, we square the discrepancies or deviations and find their mean

as an indicator of how great is their average. The squared discrepancies $(Y' - M_y)^2$ are given in column (5) of Table 69. But before finding a mean of the squared discrepancies, we weight each one for a column by the number of cases in that column. The weighted, squared discrepancy for each column will be found in the last column of Table 69. Then they are summed, and we divide by N , which is 150 in this problem, to find $\sigma_{y'}^2$, which is 110.2997. The square root of this is 10.50, which is the σ of the discrepancies

TABLE 69—THE COMPUTATION OF A CORRELATION RATIO FOR THE REGRESSION OF TIME SCORE ON CHRONOLOGICAL AGE

(1) X CA	(2) n_c	(3) Y' Time	(4) $Y' - M_y$	(5) $(Y' - M_y)^2$	(6) $n_c(Y' - M_y)^2$
14	10	11 0	-12 0	144 00	1,440 00
13	15	14 0	- 9 0	81 00	1,215 00
12	12	14 5	- 8 5	72 25	867 00
11	19	16 0	- 7 0	49 00	931 00
10	18	18 1	- 4 9	24 01	432 18
9	21	20 8	- 2 2	4 84	101 64
8	18	25 1	+ 2.1	4 41	79 38
7	15	31 3	+ 8 3	68.89	1,033 35
6	13	40 5	+17 5	306 25	3,981 25
5	9	49 8	+26 8	718 24	6,464 16
Sum ..	150		.		16,544 96 110 2997 10 50
					$\Sigma n_c(Y' - M_y)^2$ $\sigma_{y'}^2$ $\sigma_{y'}$

Remember that these are *not* the discrepancies of the observed points from the predicted Y values, for the larger these are the *lower* our correlation. We are here interested in the size of discrepancies between predicted Y values and the mean of all Y values, and the *larger* these are the *higher* our correlation. When the correlation is perfect, $\sigma_{y'}$ is as large as σ_y , for then the ratio $\sigma_{y'}/\sigma_y$ equals 1.00. When $\sigma_{y'} = 0$, the ratio equals zero. In this problem, $\sigma_y = 12.535$. The correlation ratio is therefore

$$\eta_{yx} = \frac{10.50}{12.535} = .838$$

The steps in computing a correlation ratio may be summarized as follows. Remember that for finding η_{xy} , we are dealing with *rows*

rather than columns, and so the steps will be the same except for the substitution of the word *row* for the word *column* in what follows and the substitution of X for Y .

- Step 1. Determine the mean of all the Y values and also their standard deviation.
- Step 2. Determine the means of the columns (Y').
- Step 3. Determine the discrepancies between Y' and M_y .
- Step 4. Square the discrepancies.
- Step 5. Multiply each squared discrepancy by the number of the cases in the column (n_c).
- Step 6. Sum the weighted, squared discrepancies, and divide by N . This gives $\sigma_{y'}$. From this, find σ_y .
- Step 7. Solve the ratio $\sigma_{y'}/\sigma_y$, which is η_{yx} .

The Standard Error of a Correlation Ratio.—The reliability of a correlation ratio, like the reliability of r , is given by its standard error, and this is derived by a similar formula

$$\sigma_\eta = \frac{1 - \eta^2}{\sqrt{N - 1}} \quad (75)$$

The standard error of the eta coefficient that we have just obtained is .025. The amount of correlation is therefore rather close to the true or population correlation.

The Standard Error of Estimate in a Non-linear Regression.—The standard error of estimate here can be computed as from a Pearson r [see formulas (67a) and (67b), page 216], but it can also be obtained from the knowledge that

$$\sigma_{yx}^2 + \sigma_{y'}^2 = \sigma_y^2$$

That is, the total variance in the Y distribution is made up of two components, the variance predictable from X (this is $\sigma_{y'}^2$) and the variance not predictable from X (which is σ_{yx}^2). Transposing, we have

$$\sigma_{yx}^2 = \sigma_y^2 - \sigma_{y'}^2$$

In solving for an eta coefficient, we must know both the terms on the right of this equation. For our illustrative problem, they are 157.262 and 110.2997, respectively. The difference is 46.8265, which is the nonpredicted variance. The square root of this, which is 6.84, gives us σ_{yx} .

The Relation of the Correlation Ratio to Analysis of Variance.—Those who have read the latter part of Ch VIII will find much in

common between the substance there and that of the last paragraph. We really have here an analysis of the total variance in Y into the two components that also are pertinent in the fundamental analysis-of-variance problem. The variance $\sigma^2_{y'}$ is the variance *between* means of sets (columns), and the variance σ^2_{yx} is the variance *within* sets and is separated from the other variance. Setting up a table for the comparison of the two variances as in Ch. VIII, we have Table 70. The F

TABLE 70—AN ANALYSIS OF VARIANCE BASED UPON STATISTICS DERIVED IN THE SOLUTION OF A CORRELATION RATIO

Component	Degrees of freedom	Sums of squares	Variances
Between sets	9	16,544 96	1,838 33
Within sets	140	7,023 97	50 17
Total	149	23,568 93	

$$F = \frac{1838.33}{50.17} = 36.6$$

ratio of *between* variance to *within* variance is equal to the very high value of 36.6, which is well above the very significant minimum, as we should have expected from the high correlation ratio itself. It can be seen, then, that in the simple two-variable problem, we have the two alternative methods of determining whether a significant law of relationship exists between Y and X . The only difference is that in the analysis-of-variance approach we take account of degrees of freedom and make a more stringent test of significance. Peters and Van Voorhis have urged that Kelley's "correction for bias" in eta be used to meet this situation.¹

A Test of Linearity of Regression.—Often the curvature in regression is so slight that we do not know but that it is merely a chance deviation from linearity. We therefore want some statistical test to show whether or not the curvature is probably real. Several tests of non-linearity have been proposed. Probably the most dependable one is that suggested by Fisher.² This method depends upon the already familiar chi square (see Ch. IX). For the solution of chi square here, we need to know the Pearson r for the same data for which an eta coefficient has been computed. The formula for chi square is

¹ Peters and Van Voorhis, *op. cit.*, p. 321.

² Peters and Van Voorhis, *op. cit.*, p. 319.

$$\chi^2 = (N - k) \left(\frac{\eta^2 - r^2}{1 - \eta^2} \right) \quad (77)$$

where k = number of columns (or rows).

For the problem in recent paragraphs, the Pearson r was found to be .763. By formula (77), we have

$$\begin{aligned} \chi^2 &= (150 - 10) \frac{.702224 - .582167}{1 - .702224} \\ &= (140) \frac{.120057}{.297776} \\ &= 56.4 \end{aligned}$$

With a chi square of this size and $k - 2$ degrees of freedom, the divergence between η_{yx} and r_{yx} is so great as to leave no doubt about non-linearity. So large a divergence between the two could hardly have happened if the true regression were linear.

THE BISERIAL COEFFICIENT OF CORRELATION

The biserial r is especially designed to take care of the situation in which both of the variables correlated are really continuously measurable, but one of the two is for some reason reduced to two categories. This reduction to two categories may be a consequence of the only way in which the data can be obtained, as, for example, when one variable is whether or not a student passes or fails to pass a certain test item. We can well assume a continuum of ability along which individuals differ with respect to the ability required to pass this item. Those having a degree of ability above a certain crucial point do pass it, and those having a degree of ability below that crucial point fail to pass. We cannot measure amounts of this ability along that continuum with one item only, but we do know which individuals are above and which are below the division point. It is as if our grouping were so coarse in this variable as to be confined to two class intervals rather than a dozen or so. If we are prepared to justify normality of distribution in this dichotomous variable, we have a formula by which a coefficient of correlation can be computed.

Computation of a Biserial r .—The principle upon which the formula for a biserial r is based is that with zero correlation there would be no difference between means, and the larger the difference between means the larger the correlation. The general formula for biserial r is

$$r_{bi} = \frac{M_p - M_q}{\sigma_t} \times \frac{pq}{y} \quad (78)$$

where M_p = mean of the higher group in the dichotomous variable, for example, the one having more of the ability in which the sample is divided into two subgroups.

M_q = mean of the lower group.

p = proportion of the cases in the higher group.

q = proportion of the cases in the lower group.

y = ordinate of the normal distribution curve with surface equal to 1.00, at the point of division between segments containing p and q proportions of the cases (see Fig. 40).

σ_t = standard deviation of the total sample in the continuously measured variable.

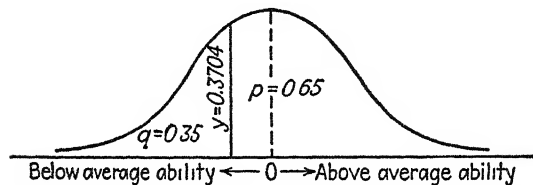


FIG. 40.—A normal distribution of the cases along the scale of ability to pass the test item in question. The area to the right of the ordinate shown represents the 65 per cent who passed the item and the area to the left represents the 35 per cent who failed to pass.

Table 71 presents typical data for the correlation between an item and total test scores. The passing group were distributed as shown; also, the failing group. The proportions passing and failing are .65

TABLE 71—DISTRIBUTIONS OF SCORES FOR TWO GROUPS OF STUDENTS—THOSE PASSING AND THOSE FAILING IN RESPONDING TO A CERTAIN TEST ITEM, AND A COMBINED DISTRIBUTION

	Scores											n/N
	40-49	50-59	60-69	70-79	80-89	90-99	100-109	110-119	120-129	130-139	n	
Passing students		1	3	10	27	30	26	21	7	5	130	65 = p
Failing students	2	6	4	11	21	16	7	3		.	70	35 = q
Total	2	7	7	21	48	46	33	24	7	5	200	1 00

and .35, respectively. The y ordinate (from Table C) is .3704. The distribution of the total group is assumed to be as indicated in Fig. 40. The computation of the biserial r proceeds as follows:

$$\begin{aligned}
 r_{bi} &= \frac{98.27 - 83.64}{17.68} \times \frac{(.65)(.35)}{.3704} \\
 &= \frac{3.328325}{6.548672} \\
 &= .508
 \end{aligned}$$

The Standard Error of r_{bi} .—The standard error of a biserial r is given by the formula

$$\sigma_{r_{bi}} = \frac{\frac{\sqrt{pq}}{y} - r^2}{\sqrt{N}} \quad (79)$$

where the symbols have already been defined above.
In this problem

$$\begin{aligned}
 \sigma_{r_{bi}} &= \frac{\frac{.4770}{.3704} - .258064}{\sqrt{200}} \\
 &= \frac{1.0297}{14.142} \\
 &= .073
 \end{aligned}$$

This standard error may be interpreted as usual, and we find that the obtained r_{bi} is so large as undoubtedly not to be arising from an uncorrelated population.

Another Formula for Biserial r .—More recently a more convenient formula has come into use for the biserial r .¹ It is

$$r_{bi} = \frac{M_p - M_t}{\sigma_t} \times \frac{p}{y} \quad (80)$$

where the only new symbol $= M_t$, the mean of the total sample. The greater convenience of this formula over the other is that we might as well compute M_t while we are computing σ_t , and so formula (80) gives us one less distribution to deal with. A good type of work sheet for solution by this formula is shown in Table 72. It is convenient to use the same guessed mean for both the component distribution and for the total distribution. By this procedure, the biserial r and its σ_r come out the same as we have already seen.

A Caution.—It should be emphasized that the dichotomous variable must be normally and continuously distributed, or the computed r will

¹ Dunlap, J. W., Note on computation of biserial correlations in item evaluation. *Psychom.*, 1936, 1 (June), 51-60.

TABLE 72.—SOLUTION OF MEANS AND STANDARD DEVIATION NECESSARY FOR THE COMPUTATION OF A BISERIAL r

Scores	x'	f_p	$f_p x'$	f_t	$f_t x'$	$f_t x'^2$
130-139	+4	5	+20	5	+20	80
120-129	+3	7	+21	7	+21	63
110-119	+2	21	+42	24	+48	96
100-109	+1	26	+26	33	+33	33
90- 99	0	30	0	46	0	0
80- 89	-1	27	-27	48	-48	48
70- 79	-2	10	-20	21	-42	84
60- 69	-3	3	- 9	7	-21	63
50- 59	-4	1	- 4	7	-28	112
40- 49	-5			2	-10	50
Sums..	130	+49	200	-27	629

$$\begin{aligned}
 c'_p &= +.377 & c'_t &= -.135 & \sigma_t &= 10 \sqrt{.629/200 - .135^2} \\
 c_p &= +3.77 & c_t &= -1.35 & &= 10 \sqrt{3.1268} \\
 \text{Means } M_p &= 98.27 & M_t &= 93.15 & &= 17.68
 \end{aligned}$$

have little meaning and in some instances will be absurd. The writer has encountered some biserial r 's greater than 1.00 when the method was wrongly applied. Also to be noted is the fact that the dichotomy should not be too lopsided; probably never more uneven than a .9 to .1 division.

TETRACHORIC CORRELATION

Perhaps the most difficult coefficient to compute by formula is the tetrachoric r ; yet because it is coming into such general use the method will be described here. This coefficient is designed for use when distributions for both variables are reduced artificially to two categories each. *It assumes actually continuous and normally distributed variables.* In other cases, its use is questionable if not absurd.

Assumptions Underlying the Tetrachoric r .—A problem in which the tetrachoric r may be computed is illustrated in Table 73, if we are willing to make the necessary assumptions. These data represent the numbers of students responding "Yes" and "No" to two questions in a personality questionnaire. Question I was, "Do you enjoy getting acquainted with most people?" and Question II was, "Do you prefer to work with others rather than alone?" Out of 930 replies to both questions, we have the numbers who responded similarly (cells a and d in Table 73) and the number who responded differently to the two

questions (cells *b* and *c*). It is obvious that in the case of a perfect positive correlation, all the cases would fall in cells *a* and *d*. In a perfect negative correlation, they would fall in cells *b* and *c*. In a zero correlation, the frequencies would be proportionately distributed in the four cells.

TABLE 73.—FOURFOLD TABLE FROM WHICH A TETRACHORIC COEFFICIENT OF CORRELATION IS COMPUTED

	Question I				Division ordinate	Point deviate
	Yes	No	Total	Proportion		
Question II	Yes	374 (<i>a</i>)	167 (<i>b</i>)	541 (<i>p</i>)	.3905 (<i>y</i>)	2070 (<i>z</i>)
	No	186 (<i>c</i>)	203 (<i>d</i>)	389 (<i>q</i>)		
	Total	560	370	930	1 000	
	Proportion	.602 (<i>p'</i>)	398 (<i>q'</i>)	1 000		
	Ordinate	.3858 (<i>y'</i>)				
	Deviate	2585 (<i>z'</i>)				

The assumption of continuity and normality of distribution can be defended as follows: It is unlikely that all who respond "Yes" to either question do so with equal degree of affirmation. It is similarly unlikely that those who respond "No" do so with equal degree of negation. It is most likely that either question represents a continuum of behavior extending from strong affirmation at the one extreme to strong negation at the other. Continuity is thus the probable state of affairs, not a discrete dichotomy. If a continuum is granted, the general law of unimodal distribution approaching normality in psychological traits may be cited in defense of the other requirement. By making the necessary assumptions, at any rate, many things can be done with such data that would otherwise be impossible. As in most statistical operations where true form of distribution is unknown, we can here remember that we have taken the chance of faulty assumptions and interpret results with the requisite reservation.

The Equation for the Tetrachoric *r*.—The complete equation for the tetrachoric *r* is indeed a long and complicated one, involving a series including many of powers of *r*. The first few terms included, it reads

$$r_t + r_t^2 \frac{zz'}{2} + r_t^3 \frac{(z^2 - 1)(z'^2 - 1)}{6} + \dots = \frac{ad - bc}{yy'N^2} \quad (81)$$

The symbols will be explained with reference to Table 73. The letters a , b , c , and d refer to the frequencies in the four cells of the fourfold table. r_t is given the subscript to indicate that it is a tetrachoric r . Numerically, it is equivalent to a Pearson r .

In Table 73, it will be noted that the distribution of responses to Question I is given in terms of proportions p' and q' . The distribution of all responses to Question II is similarly given in terms of p and q . These proportions are required for finding the values for the y 's and z 's in formula (81). The symbols z and z' stand for the standard measurements on the base line of the normal distribution curve at the points of division of cases in the two distributions.

From Table C, we find that z is .2070 and z' is .2585. The symbols y and y' stand for the ordinates in the normal curve at the points of division. From Table C, we find that they are .3905 and .3858, respectively. N is, of course, 930. We now have all the values except r_t , for which we must solve the equation.

The Solution for a Tetrachoric r .—Let us ignore all terms involving higher powers of r_t than the second. We can then reduce formula (81) to a quadratic equation that is readily soluble but that will give only an approximation to r_t . The terms ignored are rather small, however, and so can be disregarded. With substitutions, the equation becomes

$$r_t + \frac{(.2070)(.2585)}{2} r^2 = \frac{(374)(203) - (186)(167)}{(.3905)(.3858)(930^2)}$$

which reduces to

$$r_t + .026755r^2 = .344279$$

It is well in this solution to carry at least six decimal places in order to assure a sufficient number of non-zero digits later.

We now have arrived at a quadratic equation, which, with rearrangement of terms becomes

$$.026755r^2 + r - .344279 = 0$$

And this is in a form to which we can readily apply a standard algebraic solution. If the standard quadratic equation is written

$$ar^2 + br + c = 0$$

we can solve for r by using the following formula:

$$r = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \quad (82)$$

In our own equation for r above, a is .026755, b is 1.0, and c is $-.344279$.

Substituting these in formula (82), we have

$$\begin{aligned} r_t &= \frac{-1.0 \pm \sqrt{1 - 4(.026755)(-.344279)}}{2(.026755)} \\ &= \frac{-1.0 \pm 1.01825}{.05351} \end{aligned}$$

Here we have a choice between the positive or the negative square root. If we choose the negative one, the numerator becomes -2.01825 , which would give us an r_t equal to about -40 . This is obviously absurd, since no law-abiding r can exceed 1. Taking the positive root, we have

$$\begin{aligned} r_t &= \frac{.01825}{.0535} \\ &= .341, \text{ or } +.34 \end{aligned}$$

We can now say that, if our assumptions about the two distributions are granted, the correlation between an expressed enjoyment of getting acquainted with people and an expressed preference for working with others is $+.34$. For greater refinement in the solution, we could now treat $.341$ as a trial value for r_t in equation (81) and see how much discrepancy is involved when the term having r^3 in it is included in the calculations. We could make any change in r_t that seemed necessary for a better satisfaction of the equation and by successive trial-and-error maneuvers arrive at a more exact choice of r_t . Probably most data are not of sufficient number or precision to justify the extra labor involved in this. The discrepancy involved when all powers higher than two are ignored in equation (81) is probably much smaller than the standard error of r_t .

The Standard Error of a Tetrachoric r .—The tetrachoric r is less reliable than the Pearson r , being as much as 50 per cent more variable. r_t is most reliable (1) when N is large, as is true of all statistics, (2) when r is large, as is true of other r 's, but also (3), when the divisions into two categories are close to the medians. The complete formula for σ_{r_t} is entirely too long to be practical; so it will not be given here. But when $r = 0$, the formula is much simpler and reads

$$\sigma_{r_t} = \frac{\sqrt{pp'qq'}}{yy' \sqrt{N}} \quad (83)$$

where the symbols mean the same as in formula (81) or in Table 73. This is the most useful of the formulas for σ_{r_t} , at any rate, because it permits testing the null hypothesis. If the correlation in the population

is zero, samples of the size we obtained would yield r 's with a standard error as given by this equation. For the 930 cases in our problem,

$$\begin{aligned}\sigma_{r_t} &= \frac{\sqrt{(.582)(.602)(.418)(.398)}}{(.3905)(.3858) \sqrt{930}} \\ &= .053\end{aligned}$$

Since the obtained r_t is .34, being more than 2.6 times this standard error, we can be quite positive that the two qualities represented by the two questions are really correlated in the population. Had this been a Pearson r , σ_r (for $r = 0$) would have been .033. This gives a rough idea of the relative degree of sampling fluctuation of the two kinds of r and also bears out the statement made earlier that the tetrachoric r is about 50 per cent more variable. This fact should impress one with the importance of using a larger sample when r_t is to be the index of correlation. Roughly, to attain the same degree of reliability in a tetrachoric r , one needs more than twice the number of cases as in the use of the Pearson r . For very dependable results, it is recommended that N be at least 200 and preferably 300 when r_t is to be computed.

Reducing Distributions in Class Intervals to Fourfold Tables.—

Data need not be obtained in two categories each way in order to apply the tetrachoric solution for r . Any scatter diagram, in fact, can be reduced to two groups each way by making arbitrary divisions. Such a division should be made as nearly as possible at or near the median in each distribution. Table 74 shows a scatter diagram in which reduction to a fourfold table would be highly desirable. A Pearson r computed with so few class intervals each way would be highly influenced by the error of grouping (see page 253). The very large number of cases renders the reduction in reliability of r by computing r_t of small importance. The divisions suggested in Table 74 come between the B 's and C 's for distribution of school marks and at an IQ of 90 for intelligence rating. The revised correlation distribution is seen in Table 74.

The Graphic Solution of Tetrachoric r .—When a large number of tetrachoric r 's must be computed, considerable saving of labor is provided by the Thurstone computing diagrams.¹ These are highly recommended and they yield two-place accuracy with little effort after the fourfold table is completely reduced to the status of proportions throughout, as in Table 74. From the computing diagrams

¹ Chesire, L., Saffir M. and Thurstone, L. L., Computing diagrams for the tetrachoric correlation coefficient. Chicago: University of Chicago Bookstore, 1938.

r_t for these data is estimated to be $+.79$. The correlation of the two questions previously cited is estimated from the diagrams to be $+.34$, which checks exactly with the arithmetical solution.

TABLE 74—THE REDUCTION OF A SCATTER DIAGRAM TO A FOURFOLD TABLE PREPARATORY TO THE COMPUTATION OF A TETRACHORIC COEFFICIENT OF CORRELATION^{*}
Mark in Schoolwork

<i>IQ</i>	F	D	C	B	A	Total
120 and above			12	32	40	84
110-119		4	23	66	23	116
100-109	1	10	67	77	15	170
90- 99	1	22	133	40	3	199
80- 89	8	71	125	21	2	227
70- 79	36	92	24	1	.	153
Below 70	27	36	4			67
Total	73	235	388	237	83	1,016

<i>IQ</i>	In terms of frequencies			In terms of proportions		
	C, or below	A or B	Total	C, or below	A or B	Total
90 or above	273	296	569	269	291	560
Below 90 ..	423	24	447	416	024	440
Total	696	320	1016	685	315	1 000

* Adapted from Cobb, M V. The limits set to educational achievement by limited intelligence. *J. educ Psychol*, 13, 1922, p 449. By permission of the publisher.

THE PHI COEFFICIENT

When the two distributions correlated are really dichotomous, when the two classes are separated by a real gap between them and previous correlational methods do not apply, we may resort to the phi coefficient. This was designed for so-called "point distributions," which implies that the two classes have two point values or merely represent some unmeasurable attribute. Such a case would be illustrated by eye color, sex, "living versus dead," and the like. The method can be applied, however, to data that are measurable on a continuous variable if we make certain allowances for that fact. It is a close relative of chi square, which is applicable to a wide variety of situations.

The Computation of Phi.—To illustrate the use of phi (ϕ), we shall use again some data that were previously employed with chi square (see Table 52, page 170). They are repeated here as we need them, in proportion form, in Table 75.

TABLE 75—A TABLE TO ILLUSTRATE THE CORRELATION OF ATTRIBUTES

	Normal	Feeble-minded	Both
Married	269 (α)	204 (β)	473 (p)
Unmarried	231 (γ)	296 (δ)	527 (q)
Both	500 (p')	500 (q')	1 000

The formula for the phi coefficient is

$$\phi = \frac{\alpha\delta - \beta\gamma}{\sqrt{pq p' q'}} \quad (84)$$

where the symbols correspond to the labeled cells in Table 75. The solution of ϕ for this table is

$$\begin{aligned} \phi &= \frac{(.269)(.296) - (.204)(.231)}{\sqrt{(.5)(.5)(.473)(.527)}} \\ &= \frac{.0325}{.2496} \\ &= .1302, \text{ or } .13 \end{aligned}$$

The Relation of Phi to Chi Square.—Phi is related to chi square by the very simple equation

$$\chi^2 = N\phi^2 \quad (85)$$

and phi is derived from chi square by the equation

$$\phi = \sqrt{\frac{\chi^2}{N}} \quad (86)$$

By formula (85)

$$\begin{aligned} \chi^2 &= (412)(.016952) \\ &= 6.98 \end{aligned}$$

This checks with the solution of chi square by other methods (see page 169).

The solution of a standard error for ϕ is too laborious to be of practical value. It is recommended that for tests of significance chi square be derived from phi and *with 1 degree of freedom* the significance of chi square be determined from Table E.

The Phi Coefficient in Continuous Distributions.—As was indicated before, the phi coefficient was designed for use with point distributions, and in these instances, the coefficient is numerically equivalent to the Pearson r . It can, however, be adapted to continuous distributions reduced to fourfold tables, such as the data on questions in Table 73. A phi computed from such a table is *not* numerically equivalent to a Pearson r but is considerably smaller. For the data on questions, a computed ϕ is equal to .215, which is, as we expected, much smaller than the r_t of .34.

We can estimate the corresponding Pearson r from ϕ by the equation

$$r_\phi = \frac{\phi}{.637} \quad (87)$$

That is, we need only divide ϕ by the constant .637. If we do this with our ϕ of .215, we obtain $r_\phi = .338$, or to two places, .34. For the data on relation of IQ to school marks, ϕ proves to be .500, which, when divided by .637, yields an estimate of r equal to .785, which is sufficiently close to the r_t for the same data (.79) to be very acceptable as a substitute. Since the ϕ coefficient is much easier to compute than is r_t , the Pearson r might well be estimated via this route. The test of the null hypothesis, when wanted, can be made, as was suggested, by making use of the chi square that corresponds to ϕ .

It might be added that when a genuinely dichotomous variable is to be correlated with a continuous one that is reduced to two categories artificially, the ϕ coefficient, or chi square, is about the only suitable approach. The coefficient under these circumstances is smaller than a Pearson r would be, and if one wished to determine what a corresponding Pearson r would be, the estimate is ϕ divided by .798 rather than by .637, as when both variables were continuous. But the meaning of such a figure is considerably in doubt, and its interpretation should be made only in the full light of the steps by which it was derived.

The Special Case of Phi When One Distribution Is Evenly Divided.

When one of the distributions, let us say the one for which we use p' and q' as total proportions, is evenly divided so that $p' = q' = .50$, the solution of ϕ is considerably simplified. The formula reads

$$\phi = \frac{\alpha - \beta}{\sqrt{pq}} \quad (88)$$

Applied to the data on marital status

$$\begin{aligned}\phi &= \frac{.269 - .204}{\sqrt{(.527)(.473)}} \\ &= \frac{.065}{.4993} \\ &= .130\end{aligned}$$

This particular case is useful in many an experimental situation where two separated groups are selected with equal numbers of cases. There is some question here, of course, as to how well the samples chosen represent the larger population from which they were obtained, and so interpretations should be stated with this knowledge in mind.

SOME SPECIAL PROBLEMS IN CORRELATION

The Relativity of All Coefficients of Correlation.—It is apparent that the size of the coefficient of correlation depends to some extent upon the method of computing it, although where there are real alternatives, the discrepancies are usually very small, or we can apply certain corrections to make coefficients more comparable. What is more important, coefficients computed between the same two variables by the same procedure will vary not only from sample to sample but from population to population. If there are any really absolute correlations in the universe, all variables except the two being held constant, as was indicated before, those correlations are probably either zero or 1. With contaminating variables left in, the correlations are usually between zero and 1. It is therefore really meaningless to speak of *the* correlation between intelligence and character (if it is assumed even that we know what those variables are and have properly measured them) or even between age and height or any other common variables without at the same time specifying what kind of sample we measured.

A coefficient is always relative to the kind of population sampled and to the manner in which the measurements were made. In reporting coefficients of correlation, any writer should be very careful to state all the pertinent factors that bear upon the size of his obtained correlation coefficients, and any reader should accept interpretations only when the significant circumstances are kept in mind. A few of the more common sources of variations of size of r will be reviewed briefly in what follows.

The Variability in the Correlated Variables.—The size of r is very much dependent upon the range of ability or, in more general terms, the variability of measured values, in the correlated sample. The

greater the variability, the higher will be the correlation, everything else being equal. It should be easier to predict individual differences in scholarship in a class with *IQ*'s ranging from 50 to 150 than in a class where the range is restricted to 90 to 110. If the restriction were to a range of zero (all *IQ*'s being equal) there should be no correlation whatever—the limiting case, in which, of course, no r could be computed at all. Often we know the correlation between some predictive index, such as aptitude-test score and scholarship or some vocational criterion of success as derived from one group of individuals, but we shall be applying the same index to other groups with different ranges of ability, larger or smaller. What will be the effectiveness of predictions in the new groups? If equal standard errors of estimate are assumed in the two groups, old and new, the estimate of correlation in the new group from the known correlation in the old can be made by the formula

$$\frac{\sigma_{y_0}}{\sigma_{y_n}} = \frac{\sqrt{1 - r_{y_n x_n}^2}}{\sqrt{1 - r_{y_0 x_0}^2}} \quad (89)$$

where σ_{y_0} = standard deviation of the old (known correlation) group.

σ_{y_n} = standard deviation in the new (correlation unknown) group.

$r_{y_0 x_0}$ = correlation already known.

$r_{y_n x_n}$ = correlation we wish to know.

To apply this formula, let us assume that for a small (fictitious) group of scholarship winners in which σ_y for an achievement examination in English is 8.5, the correlation between scholastic-aptitude score and English-achievement score is .35. This hardly looks promising for the aptitude test as a prognostic index. The entire group of the students, including winners and non-winners, have a standard deviation of 12.0 in the achievement examination. What is the probable correlation for this wider group? Substituting in the formula

$$\frac{8.5}{12.0} = \frac{\sqrt{1 - r_{y_n x_n}^2}}{\sqrt{1 - .1225}}$$

Squaring both sides of this equation

$$\frac{72.25}{144} = \frac{1 - r_{y_n x_n}^2}{.8775}$$

from which r is found to be .75. Had someone decided from the known correlation that *the* correlation between this aptitude test and achievement in English is only .35, he would probably have discarded the test

as worthless for predictive purposes. But by making this adjustment for different ranges of ability, it is seen that in a more widely scattering group, validity is very high. In converse manner, a test that is validated on a sample with unusually wide dispersion gives a too optimistic picture of its validity for predicting within a group of narrower dispersion. In the use of formula (89), a change in dispersion in X can be treated in the same manner if we want to predict X from Y .

Predicting in Groups with Systematic Differences among Them.—Studies of validity of tests and examinations have frequently been faulty from a number of standpoints. The use of school marks as criteria of success in training is in itself a questionable procedure, school marks being derived as they generally are on the basis of measurements of questionable reliability and validity and contaminated with irrelevant factors. This situation alone stacks the cards against high validity coefficients for predictive indices at the start.

There is another factor working against fair tests of validity that we shall face particularly here, a factor also dependent upon the unwarranted faith in school marks as absolute and dependable measures of scholarship. This factor is the indiscriminate pooling of marks from different subjects and from different instructors and treating them as if they were of the same kind of coin. Any cursory inspection of grade distributions in a single institution of learning will show that marks are not by any means of constant value when obtained from different sources. The reader is referred to the situation in Fig. 28 (page 184) where students in an English course making the same score in a common achievement examination were assigned different marks in different sections and by different instructors, probably within the same section. If it is assumed that the comprehensive examination was a valid measure of the students' relative degree of mastery of the objectives of the course, it can be seen how much other factors must have entered into the determination of the final mark in the course.

Reference to Fig. 28 will show that there is quite a range of scores, from about 85 to 125, within which students were assigned marks all the way from F to B. Only as between marks of F and A is there rather complete lack of overlapping. Striking as this situation is, it is probably rather representative of how much lack of correlation there is between school marks and genuine achievement. Much of this is due to the fluctuation of marking ideas and ideals from instructor to instructor. This variation from set to set of marks when they are collectively correlated with other measures is bound to alter the apparent amount of correlation.

As an example, in six sections of freshmen English, *within* sections the correlation between quiz averages for the semester and a final comprehensive examination ranged from .63 to .92, with an over-all correlation *within* sections, *when intersection differences had been eliminated*, of .83. Yet when the six sections were combined, *with inter-sectional differences left in*, the correlation was reduced to .71. It was interesting to find that *between* sections the correlation was $-.17$, which means that there was a very slight tendency for sections with average lower achievement to be given a higher average quiz mark! This fact accounts for the reduction in correlation from .83 to .71 when sections were combined.¹

The Correlation of Averages.—It was stated in an earlier chapter in connection with tests of significance of differences between statistics (Ch. VIII) that the correlation between averages of samples is equal to the correlation between individual pairs of measurements. *This statement assumes random samples from a homogeneous population.* It does not apply to the kind of case just described, where obviously the correlation between sets is decidedly different from that between individuals. This is due to the heterogeneity among the sets. In the instance cited above, the inter-set correlation was decidedly lower than that between individuals, but there can be other cases in which the correlation between averages is decidedly higher than that between individuals. An example of this would be the correlation between *IQ* and salary. Correlating individuals, we should find some positive correlation, but because of great variations in salary at any single *IQ* value, the correlation might not be very high. If we divided men into sets according to vocation and correlated *average IQ* with *average* salary, the coefficient would probably be very high. This is because people of different *IQ* levels gravitate to certain occupations, and occupations as such have established characteristic salary scales. Other factors that make for individual differences in salary *within* occupations are thus minimized in importance. The sampling is biased the moment we divide groups along occupational lines.

The Correlation of Parts with Wholes.—We frequently want to correlate a part measurement, such as a part of a test battery, or a test item, with the whole of which it is a part. Since the variance of the total is in part made up of the variance of the component, that fact alone introduces some degree of positive correlation. The greater the

¹ Further discussion of "within" versus "between" correlations when groups are combined will be found in E. F. Lindquist's *Statistical analysis in educational research*. Pp. 219ff.

relative contribution to the total variance by the component, the more important is this "spurious" factor. It is possible in a particular instance that the part is totally *uncorrelated* with the remaining parts and yet will be correlated with the total. If it is negatively correlated with the remaining parts, it will be less negatively correlated with the total.

If each part contributes statistically about the same amount of variance to the total or if the part is one of a great many, so that its proportion of contribution is relatively small, we can compare correlations between parts and total with some confidence that they are compared on the same basis. But if these conditions do not obtain, we should do better to correlate each part with a composite of all other parts. When such a composite is unknown or is hard to obtain, we can still estimate the correlation by means of the formula

$$r_{pq} = \frac{r_{tp}\sigma_t - \sigma_p}{\sqrt{\sigma_t^2 + \sigma_p^2 - 2r_{tp}\sigma_t\sigma_p}} \quad (90)$$

where p = part score.

t = total score.

$q = t - p$, in other words, the total with the part excluded

Index Correlations.—This is usually called *spurious index correlation* for the reason that when indices, such as IQ , EQ , or AQ , are correlated with each other, r is markedly influenced by the fact that these ratios have in common such factors as chronological age and mental age. IQ 's from different tests are derived from the MA 's derived from the two tests each divided by the same CA . If there is a range of CA in the group correlated, this fact in itself would introduce some positive correlation.

In the writer's opinion, the term *spurious* is not to be confined to this type of situation in particular; for in a sense, all correlations are spurious to the extent that they are influenced by the conditions under which they were obtained. If one remembers what IQ 's are and interprets correlations between them accordingly, no particular falsification of the facts is in question. The important thing is that one should correlate variables in the full knowledge of how the measurements were obtained, if possible, and should report to his readers the facts needed for wise interpretation, whether it be variability of the correlated group or range of CA 's involved when IQ 's have been correlated.

The real difficulty comes when investigator or reader takes IQ 's to be some real, absolute properties of individuals, on the one hand, and

when someone not oblivious to the common *CA* factor plays it up as a fatal source of "error," on the other hand. Both should remember the relative nature of all correlation coefficients. The important thing is that the wary investigator should not attribute his results to some supposed real nature of psychological or educational phenomena when some property of statistical treatment is really responsible. Nor will the sophisticated critic fail to grant the utility of certain procedures shown to be fruitful under the circumstances of operation even when some "spurious" element has entered the picture. Errors, too, are relative matters. What is an error from the point of view of one frame of reference may be the truth when the frame of reference is changed.

Correction in r for Errors of Grouping.—If in computing a Pearson r by means of grouping data in class intervals, a small number of classes either way has been used, the computed r is lowered to some degree. In the limiting case, of two classes each way, the computed r is only about two-thirds of the r had there been no grouping. This was evident in the correction suggested in the ϕ coefficient (formula 87)

TABLE 76 — CORRECTION FACTORS FOR ERRORS OF GROUPING IN THE COMPUTATION OF PEARSON'S r WHEN DISTRIBUTIONS ARE NORMAL AND MIDPOINTS OF INTERVALS STAND FOR CASES IN THE INTERVALS

Number of intervals	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Correction factor	.816	.859	.916	.943	.960	.970	.977	.982	.985	.988	.990	.991	.992	.994

where we divided an obtained ϕ by the constant .637 to take care of the error of grouping. When the number of intervals is 10 both ways, r is about 3 per cent underestimated. For any number of classes each way, we can correct for the error of grouping by dividing r by a constant corresponding to that number of classes.

Table 76 supplies the list of constants as given by Peters and Van Voorhis.¹ The constants are given for variables Y and X separately, for they frequently have different numbers of class intervals. In a correlation problem in the preceding chapter (pages 208 ff.), the obtained r was $+.61$. The number of class intervals in Y was nine and in X was eight. For nine and eight intervals, we see in Table 76 that the correction factors are .982 and .977, respectively. The r corrected for errors of grouping is therefore

¹ Peters and Van Voorhis, *op cit*, p. 398.

$$r_c = \frac{.61}{.936} = .652, \text{ or } .64$$

The general formula is

$$r_c = \frac{r}{c_x c_y} \quad (91)$$

where c_x and c_y = correction factors for variables X and Y (derived from Table 76).

When the number of intervals in either X or Y is less than 10, it is probably well to use this correction process, unless N is so small that r is highly unreliable. It would be of most importance to use it when r is near the borderline of significance, as determined by the t ratio. The correction factors given in Table 76 apply only for the artificial grouping procedure where the midpoint of the interval is used as the index value for that interval and where distributions are normal. For other, less common instances, see the reference below.¹

Exercises

- 1 Compute by the rank-difference method the correlation between the first 20 scores in ascendance-submission and in masculinity-femininity in Data *AA* (page 225). Find the standard error of rho. Interpret your results.
- 2 Compute for Data *Z* (page 197) a correlation ratio for the prediction of Y from X . Find the standard error of eta and the standard error of the estimate. Apply the chi-square test of linearity. Interpret your results.
- 3 Find from the literature three applications of the correlation ratio. State how the author used eta, and give his reasons, if stated. What subsidiary tests (of linearity, etc.) were made? Make your judgment as to the effectiveness of the uses of eta in the cases cited.
- 4 If you have mastered the analysis-of-variance procedures as described in Ch. VIII, make the application as suggested in this chapter to Data *Z*, following your solution of the correlation ratio.
- 5 In the data in Table 74 (page 245), combine the distributions receiving marks of A, B, or C into a single composite, also, in another composite, combine those receiving marks of D and F. Compute for these data a biserial r between scores and marks. Find the standard error of r_b . Interpret your results.
- 6 Compute a tetrachoric coefficient of correlation for Data *X* (page 196). Determine whether or not the correlation is probably significant. If the Thurstone computing diagrams are available, check your solution by this means.
- 7 Cite some fourfold tables found in this book to which the tetrachoric correlation method should be applied, and cite some others to which it should not be applied.
- 8 Reduce to a fourfold table preparatory to computing a tetrachoric r the scatter diagram given in Data *Z*. Do the same for Data *V* (page 174) and Data *BB* (page 226).

¹Peters and Van Voorhis, *op cit*, p. 398.

9 Find in this volume, or in any other source, data to which the phi-coefficient method of correlation may properly be applied. Give reasons.

10 Compute a phi coefficient for Data X (page 196), and make the necessary correction to yield an estimate of the Pearson r . If Exercise 6 has been completed, compare with the r_t found there.

11 Find in the literature examples of coefficients of correlation that might be regarded as spurious from some points of view. How did the author interpret them? How would you interpret them?

12 Apply the correction-for-grouping process to some product-moment coefficient you have obtained or to one you find uncorrected in the literature.

13 Compute a Pearson r for the data in Table 74 (page 245), and correct it for errors of grouping. How does the change in the corrected r compare with σ_r ? How do the uncorrected and corrected Pearson r 's compare with the tetrachoric r given for the same data?

CHAPTER XIII

MULTIPLE AND PARTIAL CORRELATION

MULTIPLE CORRELATION

Independent and Dependent Variables.—Thus far we have been dealing with correlations between two things at a time and the prediction of some variable Y from another variable X , or vice versa. Actual relationships between measured things in psychology and education are by no means so simple as that. One variable is found associated with, or dependent upon, more than one other variable at the same time. When we can think of some variables as being causes of another one, or even when we merely want to predict that one from our knowledge of several others that are correlated with it, we call the one variable the *dependent* variable and the ones upon which it depends the *independent* variables. The independent variables are so called because we can manipulate them at will or because they vary by the nature of things, and in consequence, we expect the dependent variable to vary accordingly.

Whether or not a certain color is liked depends upon several factors: its hue (whether yellow, red, or purple, etc.), its lightness (whether light, medium, or dark), and its chroma (saturation or density). The affective value of the color also depends upon its area, its use, and its background. We are here naming independent variables upon which the affective value of a color depends. Insofar as each one is a determiner of agreeableness of color, it will exhibit some correlation individually with affective value. The size of any one of these correlations will depend upon the relative strength of that factor and also upon how well the other factors have been neutralized, as they should be in a good experimental situation.

The Coefficient of Multiple Correlation.—When we are interested in the amount of correlation between a dependent variable and two or more others simultaneously, we are dealing with a multiple-correlation problem. The multiple coefficient of correlation indicates the strength of relationship between one variable and two or more others taken simultaneously. The multiple correlation is not merely the sum of the correlations of the dependent variable and the various independent

variables taken separately. Obviously, there would be instances in which these would add up to more than 1.00. The reason is that independent variables themselves are usually overlapping (intercorrelated) and so duplicate one another to some extent. In this we see one important principle of multiple correlation. The multiple R is related to the intercorrelation of independent variables as well as to their correlation with the dependent variable. The interdependence of the factors suggested for affective value of colors is probably not so apparent as in the case of factors related to achievement in college algebra. Here we can think of such predictive factors as intelligence and high-school marks, which being related duplicate one another to some extent in predicting achievement in college algebra. Hours of study and interest also bear much in common and so are not completely independent determiners of success in algebra.

A Multiple-correlation Problem.—In Table 77 are presented some data that call for the multiple-correlation solution. Four of the

TABLE 77 —INTERCORRELATIONS AMONG FIVE VARIABLES, INCLUDING ONE INDEX OF SCHOLARSHIP AND FOUR PREDICTIVE INDICES ($N = 174$)*

Variable	X_2	X_3	X_4	X_5	X_1
X_2		.562	401	197	465
X_3	562		396	.216	583
X_4	401	396		345	546
X_5	197	215	345		365
X_1	465	583	546	365	
M_x	19 7	49 5	61 1	29 7	73 8
σ_x	5 2	17 0	19 4	3 7	9 1

X_2 = arithmetic test in the Ohio State Psychological Examination, Form 10.

X_3 = analogies test in the same examination.

X_4 = an average grade in high-school work

X_5 = student interest inquiry (measuring breadth of interest).

X_1 = an average grade for the first semester in university

* These data were abstracted from the *Ohio State Coll Bull* 58, by L. D. Hartson, and have been used in this chapter by permission.

variables (X_2 , X_3 , X_4 , and X_5) are all measures of things that supposedly determine success in college freshmen. X_1 is the dependent variable, or average freshman marks. It is customary to designate the dependent variable by X_1 , though some authors, less often, call it X_0 . An examination of Table 77 shows that the analogies test and high-school average mark have the highest correlation, when taken alone, with X_1 , whereas the interest score X_5 has the lowest. The

highest *intercorrelations* come between X_2 , X_3 , and X_4 . All represent abilities of one kind or another, and their correlations with X_5 (interests) are generally lower. This gives promise that the interest scores will contribute something to the prediction of college marks that will not have been already contributed by the other variables, and so it should pay to include X_5 in the battery of predictive indices. As a matter of experience in psychological and educational predictions, it has been found that it rarely pays to bring into a multiple-prediction situation more than four or five independent variables. By the time that this many are combined, they have fairly well covered what any additional one can do for us. This is partly a consequence of the fact that good human qualities tend to go together (to be intercorrelated) and partly that our predictive indices tend to remain in the area of abilities, ignoring personality factors, physical factors, and external circumstances.

The Solution of a Three-variable Problem.—We first take the simplest case of multiple correlation, that between the dependent variable and two independent variables. In the general problem given by the data in Table 77, we may ask what is the correlation between freshman marks on the one hand and the two variables analogies-test scores and high-school averages on the other. The simplest general formula for this case is

$$R^2_{1\ 23} = \frac{r^2_{12} + r^2_{13} - 2r_{12}r_{13}r_{23}}{1 - r^2_{23}} \quad (92)$$

where $R_{1\ 23}$ = multiple coefficient of correlation between X_1 and a combination of X_2 and X_3 .

r_{12} , r_{13} , and r_{23} = correlations among X_1 , X_2 , and X_3 .

Be sure to notice that this formula merely gives us R^2 , the square root of which is R . The immediate example we have set for ourselves is to find $R_{1\ 34}$ rather than $R_{1\ 23}$. To use formula (92), we need merely to substitute the subscripts 3 and 4 for 2 and 3. The solution is

$$\begin{aligned} R^2_{1\ 34} &= \frac{(.583)^2 + (.546)^2 - 2(.583)(.546)(.396)}{1 - (.396)^2} \\ &= \frac{.339889 + .298116 - .252108}{1 - .156816} \\ &= .457666 \\ R_{1\ 34} &= .677 \end{aligned}$$

The Multiple-regression Equation.—We also have here a prediction problem of estimating X_1 values from both X_3 and X_4 , and this calls for a regression equation that involves all three variables, in other words, a multiple-regression equation. From such an equa-

tion, we can predict an X_1 value for every individual. The correlation between these predicted values (X'_1) and the obtained ones (X_1) would be .677. This is another interpretation of a multiple coefficient of correlation. For the three-variable problem, the regression equation has the general form $X_1 = a + b_{12.3}X_2 + b_{13.2}X_3$. As in previous regression equations, the coefficient a is a constant and must be calculated from the data. The b coefficients serve the same purpose here as in the simple two-variable equation. The coefficient $b_{12.3}$ is the multiplying constant or weight for the X_2 values, and $b_{13.2}$ is the same for the X_3 values.

Solution of the b Coefficients.—We do not find the b coefficients directly from the correlations but do so indirectly through the so-called *beta coefficients*. They are given by the formulas

$$b_{12.3} = \left(\frac{\sigma_1}{\sigma_2} \right) \beta_{12.3} \quad (93a)$$

$$b_{13.2} = \left(\frac{\sigma_1}{\sigma_3} \right) \beta_{13.2} \quad (93b)$$

The betas, in turn, are found by the formulas

$$\beta_{12.3} = \frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2} \quad (94a)$$

and

$$\beta_{13.2} = \frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2} \quad (94b)$$

Similar equations apply, with change of subscripts, when the independent variables are X_3 and X_4 instead of X_2 and X_3 . In our example

$$\begin{aligned} \beta_{13.4} &= \frac{.583 - (.546)(.396)}{1 - (.396)^2} \\ &= \frac{.3668}{.8432} \\ &= .435 \end{aligned}$$

And

$$\begin{aligned} \beta_{14.3} &= \frac{.546 - (.583)(.396)}{1 - (.396)^2} \\ &= \frac{.3151}{.8432} \\ &= .374 \end{aligned}$$

We can now solve for the b coefficients by means of formulas (93ab):

$$b_{13.4} = \frac{9.1}{17.0} (.435) = .233$$

and

$$b_{14.3} = \frac{9.1}{19.4} (.374) = .175$$

For the complete regression equation, the a coefficient is still lacking. It is given by the formula

$$a = M_1 - b_{13.4}M_3 - b_{14.3}M_4 \quad (95)$$

Inserting the known values

$$\begin{aligned} a &= 73.8 - (.233)(49.5) - (.175)(61.1) \\ &= 73.8 - 11.53 - 10.69 \\ &= 51.58 \end{aligned}$$

The complete regression equation will then read

$$X'_1 = 51.58 + .233X_3 + .175X_4$$

To interpret the equation, we may say that for every unit increase in X_3 , X_1 is increasing .233 unit and that for every unit increase in X_4 , X_1 is increasing .175 unit. To apply the equation to a particular student whose X_3 score is 25 and whose X_4 score is 32, we predict that his X_1 score will be

$$X'_1 = 51.58 + 5.82 + 5.60 = 63.00$$

We use X'_1 to stand for his predicted average freshman mark, because he has an actual average mark that we call X_1 . Some other examples

TABLE 78—SOME PREDICTIONS OF SCHOLARSHIP MARK FROM MEASURES IN TWO VARIABLES

	Student				
	A	B	C	D	E
X_3 analogies score	25	27	48	85	87
X_4 high-school average	32	61	65	90	52
$b_{13.4}X_3$	5 82	6 29	11 18	19 80	20 27
$b_{14.3}X_4$	5 60	10 68	11 38	15 75	9 10
X'_1 (predicted mark)	63 0	68 6	74 1	87 1	81 0

of individual students are presented in Table 78 to show how various combinations of values for X_3 and X_4 point to corresponding values of X_1 .

Calculating the Multiple R from Beta Coefficients.—If the beta coefficients are known, the shortest route to the multiple R is by way of the equation

$$R^2_{1\ 23} = \beta_{12}r_{12} + \beta_{13}r_{13} \quad (96)$$

Again, note that this gives R^2 , from which the square root must be obtained. For the scholarship data and variables X_3 and X_4

$$\begin{aligned} R^2_{1\ 34} &= (.435)(.583) + (.374)(.546) \\ &= .253605 + .204204 \\ &= .457809 \\ R_{1\ 34} &= .677 \end{aligned}$$

as was found by formula (92) previously.

Interpretation of a Multiple R .—Once computed, a multiple R is subject to the same kinds of interpretation, as to size and importance, as were described for a simple r in Ch. XI. One kind of interpretation is in terms of R^2 , which we call the *coefficient of multiple determination*. This tells us the proportion of variance in X_1 that is dependent upon or predicted by X_3 and X_4 combined. In this case, R^2 is .4578, and we can say that 45.78 per cent of the variance in freshman marks is accounted for by whatever is measured by the analogies test and by high-school marks taken together, eliminating from double consideration things that they have in common. The remaining percentage of the variance, which is 54.13 ($1 - R^2$), is still to be accounted for. This remainder is given the symbol K^2 and is known as the *coefficient of multiple nondetermination*. This is consistent with the fact that R^2 and $K^2 = 1.0$, just as $r^2 + k^2 = 1.0$ in the simple correlation problem.

Relative Contribution of Independent Variables.—Since the coefficient of multiple determination, or R^2 , is composed of the two factors in formula (96) and since each factor pertains only to one of the independent variables, it is permissible to take each factor as indicating the contribution of one independent variable to the total predicted variance of X_1 . This being the case, the first factor, .253605, indicates the contribution to freshman scholarship by ability in the analogies test, and the second factor, .204204, indicates the contribution of the high-school average. Rounded, in terms of percentages, these are 25.4 and 20.4, respectively. This enables us to obtain a more definite idea of the relative importance of each variable in the regression

equation. We can say that ability in the analogies test, with what it has in common with high-school scholarship held constant, contributes about 25 per cent to freshman scholarship and that high-school marks, apart from that component related to analogies-test ability, contributes about 20 per cent. We cannot take these as final or absolute, for there are other factors contributing to freshman scholarship level that have not been similarly eliminated from consideration. But it is of much value to be able to compare contributions of variables to outcomes in this manner.

The Standard Error of Estimate from Multiple Predictions.—The standard error of estimate is again brought in to indicate about how far the predicted values would deviate from the obtained ones. The formula is the same as previously, except that the multiple R is substituted for r . It now reads

$$\sigma_{1\ 23} = \sigma_1 \sqrt{1 - R^2_{1\ 23}} \quad (97)$$

In the illustrative problem

$$\begin{aligned} \sigma_{1\ 34} &= 9.1 \sqrt{1 - .457809} \\ &= 9.1 \times .736 \\ &= 6.7 \end{aligned}$$

We can now say that two-thirds of the predicted X_1 values will lie within 6.7 points of the obtained X_1 values. The margin of error *with* knowledge of X_3 and X_4 is 73.6 per cent as great as the margin of error would be without that knowledge. These conclusions presuppose predictions made on the basis of the regression equation that was obtained, and predictions made for individuals belonging to the population and sampled at random.

The index of forecasting efficiency may also be used by way of interpretation and because of its close relation to the standard error of estimate may be mentioned at this point. The formula is the same as for a Pearson r (see page 223). In the example of our three variables, $E = 26.4$ per cent, which means that predictions by means of the equation are 26.4 per cent better than those made merely from a knowledge of the mean of the X_1 values.

The Reliability of a Multiple R .—The standard error of R is the same as for an ordinary Pearson r (see page 209), and the usual interpretations may be applied, with the same reservations. In testing for reasonableness of the null hypothesis (no correlation), Table D is again most convenient. For a certain number of degrees of freedom, the lowest significant and very significant R 's are given. The number

of degrees of freedom for the multiple-correlation problem is $N - m$, where N is the number of items correlated and m is the number of variables, in this case 3. Since N is 174, there are 171 degrees of freedom. In the table, for 150 degrees of freedom and three variables, an R of .198 is significant, and one of .244 is very significant. We have no cause to doubt some real correlation in the population sampled.

TABLE 79—SOLUTION OF A MULTIPLE-CORRELATION PROBLEM BY THE DOOLITTLE METHOD

Column number		2	3	4	5	1	Check
Variable		X_2	X_3	X_4	X_5	X_1	Sum
Row	Instruction						
A	r_{2k}	1 0000	5620	.4010	1970	4560	2 6250
B	$A \div (-A2)$	-1 0000	-.5620	-.4010	-.1970	-.4560	-2 6250
C	r_{3k}		1 0000	3960	2150	5830	2.7560
D	$A \times B3$		- 3158	- 2254	- 1107	- 2613	-1 4752
E	$C + D$		6842	1706	1043	.3217	1 2808
F	$E \div (-E3)$		-1 0000	-.2493	-.1524	-.4702	-1 8720
G	r_{4k}			1 0000	3450	5460	2 6880
H	$A \times B7$			- 1608	- 0790	- 1865	-1 0526
I	$E \times F4$			- 0425	- 0260	- 0802	- 3193
J	$G + H + I$			7967	2400	2793	1 3161
K	$J \div (-J4)$			-1 0000	-.3012	-.3506	-1 6519
L	r_{5k}				1 0000	.3650	2.1220
M	$A \times B5$				- 0388	- 0916	- 5171
N	$E \times F5$				- 0159	- .0490	- 1952
O	$J \times K5$				- 0723	- 0841	- 3964
P	$L + M + N + O$				8730	1403	1 0133
Q	$P \div (-P5)$				-1 0000	-.1607	1 1607

Multiple Correlation with More than Three Variables.—With more than three variables, the best solution of a regression equation and of a multiple R is by means of the Doolittle method. This procedure will be outlined step by step for a five-variable problem. We shall use all the variables represented in Table 77, asking what regression weights would best predict X_1 from the other four simultaneously.

and what the correlation of those predictions with obtained X_1 values would be.

First we prepare a work sheet like that in Table 79. There is a column for every variable and the numbering corresponds. A last column is introduced for the purpose of checking the calculations, as will be explained. The rows are designated by letters, and in the first column, a shorthand instruction is noted. These will be explained.

- Step 1. Record in row *A* the correlations with X_2 . These are obtained here from Table 77. In column (2), a coefficient of 1.0000 is inserted, because it is demanded by the Doolittle method. We are going to carry four decimal places throughout the solution (one more than those given in the r 's), so we record all numbers to four places.
- Step 2. Sum the values recorded in row *A*, and give the sum in the last or "check" column. This will be used later.
- Step 3. Divide the numbers in row *A* each by -1.0000 . In the table, the instruction reads " $A \div (-A2)$," which means that each number in row *A* is to be divided by the number that appears at $A2$ [row *A*, column (2)] with sign changed. This includes the last column as well.
- Step 4. Record in row *C* all the remaining correlations with X_3 . We say "remaining," because one is already recorded, namely, r_{23} . The value of 1.0000 is recorded at $C3$.
- Step 5. Sum all the correlations with X_3 , including the .5620 in row *A*. Record the sum in the "check" column.
- Step 6. The numbers in row *D* are found by the instruction " $A \times B3$," which means to multiply all the numbers in row *A* [beginning in column (3)] by the number that appears in row *B* and column (3). This number is $-.5620$ in Table 79.
- Step 7. Row *E* calls for the addition of all numbers in rows *C* and *D*.
- Step 8. Row *F* calls for the division of all numbers in row *E* by the number appearing in row *E* and column (3), with sign changed. This number, with sign changed is $-.6842$.
- Step 9. We are ready for the first checking of calculations. Sum the values in row *F*, *not* including the last column. This should equal approximately -1.8720 in this particular problem, which was found by the steps already described. If there is a serious discrepancy here (other than in the fourth decimal place), check row *E* by adding values up to the check column. If this does not check, there is an error further

back, and some recalculating is in order. All checks should be satisfied before proceeding.

- Step 10. In row *G*, record remaining correlations with X_4 , with 1.0000 at *G4*.
- Step 11. Sum *all* the correlations with X_4 , and record in the last column in row *G*.
- Step 12. Values in row *H* are the products of values in row *A* times the number at *B4*. This number is $- .4010$.
- Step 13. Values in row *I* are the products of numbers in row *E* times the number at *F4*, which is $- .2493$.
- Step 14. Sum the numbers in rows *G*, *H*, and *I* for each column.
- Step 15. Divide row *J* through by the number at *J4*, with sign changed; in other words, by $- .7967$.
- Step 16. Check by summing row *K* up to the last column. Does the sum agree with the number already found in that column?
- Step 17 and after. By now the abbreviated instructions for each row should be clear by analogy to those already given. The final check is made in column (*Q*).

The illustrative solution is set up for a five-variable problem, but a larger number of variables would be treated in a similar manner simply by extending the table to more rows and columns. A smaller number of variables would mean fewer rows and columns. It will be noticed that the table is set up in terms of *blocks* of work, each one beginning with the entrance of correlations for a new variable and ending by dividing by a number that will assure a -1.0000 as the first number in the last row of that block. The work will be found to be very systematic throughout. Any variable may be treated as the dependent variable, but it must then occupy the next to the last column in the table.

Solution of the Beta Coefficients.—The work represented in Table 79 is only a part of the Doolittle solution. The end result gives the beta coefficients, which we find by what is called a “back solution,” so called because we work in a backward direction, as compared with the work in Table 79. This work can be tabulated, but it is probably clearest to the beginner in the form of equations. The first beta found is β_{15} , which can be located without further ado in Table 79. It is the number at the intersection of row *Q* and column (1), but with sign changed (in other words, it is described as $-Q1$). β_{15} is therefore $+ .1607$. The other betas require more work; so we will follow the procedure step by step, including again the first step already taken, for the sake of completeness.

$$\text{Step 1. } \beta_{15} = -Q_1 = +.1607.$$

$$\text{Step 2. } \beta_{14} = -K_1 + \beta_{15}(K_5) = .3506 + (.1607)(-.3012) \\ = +.3022.$$

$$\text{Step 3 } \beta_{13} = -F_1 + \beta_{15}(F_5) + \beta_{14}(F_4) \\ = .4702 + (.1607)(-.1524) + (.3022)(-.2493) \\ = +.3703.$$

$$\text{Step 4. } \beta_{12} = -B_1 + \beta_{15}(B_5) + \beta_{14}(B_4) + \beta_{13}(B_3) \\ = .4650 + (.1607)(-.1970) + (.3022)(-.4010) \\ + (.3703)(-.5620) \\ = +.1039.$$

Before going further, it is well to check the calculations of the beta coefficients. This can be done by the use of the equation

$$\beta_{12}r_{25} + \beta_{13}r_{35} + \beta_{14}r_{45} + \beta_{15} = r_{15}$$

Substituting known values

$$(.1039)(.197) + (.3703)(.215) + (.3022)(.345) + .1607 = .3651$$

Since $r_{15} = .365$, the check is satisfied, and we may assume that there has been no error in computing the betas. This checking procedure can be summarized as in Table 80, which provides a convenient work plan

TABLE 80—A CHECK UPON THE COMPUTATION OF THE BETA COEFFICIENTS

	β_{1k}	r_{k5}	$\beta_{1k}r_{k5}$
X_2	1039	197	0205
X_3	3703	215	.0796
X_4	3022	345	1043
X_5	1068	1 000	1607
			$\Sigma \text{ 3651} = r_{15}$

The Solution of Regression Weights and the Multiple R .—Each b coefficient needed in the multiple-regression equation is found from its corresponding beta. Equations like those in formulas (93*a*) and (93*b*) apply. The b weight for X_2 should now read in full $b_{12 \text{ } 345}$ to indicate that we are interested in the relation of X_1 to X_2 , other variables, X_3 , X_4 , and X_5 , being held constant. For the sake of brevity (as, indeed, we have already done for the betas), we shall denote the b 's only by the first two subscript numbers b_{12} , b_{13} , etc. In the solution of a multiple R , equation (96) needs to be extended to include as

many terms as there are variables. R^2 is the sum of the products of beta times its corresponding r , *i.e.*,

$$R^2 = \beta_{12}r_{12} + \beta_{13}r_{13} + \beta_{14}r_{14} + \beta_{15}r_{15} + \dots \quad (98)$$

The a coefficient in the equation is also found by formula (95), extended with as many terms as necessary. It is the mean of the X_1 values minus the products of other means times their corresponding b weights, as

$$a = M_1 - b_{12}M_2 - b_{13}M_3 - b_{14}M_4 \dots \quad (99)$$

All these operations are conveniently carried out in a work sheet like Table 81, where R and the regression weights are systematically

TABLE 81—SOLUTION OF THE REGRESSION COEFFICIENTS FOR THE
MULTIPLE-REGRESSION EQUATION

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	β_{1k}	r_{1k}	$\beta_{1k}r_{1k}$	σ_1/σ_k	b_{1k}	M_k	$(-M_k)b_{1k}$
X_2	1039	465	048314	.556	.198	49.5	-.9801
X_3	.3703	583	215885	535	.142	61.1	-.8676
X_4	3022	546	165001	469	.395	29.7	-.11732
X_5	1607	.365	058655	2.459			
			$\Sigma .487855 = R^2$				$\Sigma -.33794$
			$.698 = R$				73.800
							$a = 40.006$

calculated. The second column contains the four betas. The third contains the original or raw correlations of the four variables with X_1 . The subscript k stands for variables 2 to 5 in turn. The fourth column contains the cross products of betas times corresponding r 's. Their sum is R^2 , which here is .487855; and by taking the square root, R is .698. This R , with full subscript, would read $R_{1\ 2345}$.

So much for the multiple R , which we see is not increased very much by including two more variables (X_2 and X_5) over that obtained when we used only X_3 and X_4 . Then R equaled .677. The coefficient of determination is now .4879, or we have accounted for 48.8 per cent of the variance of freshman scholarship, as compared with 45.8 per cent without using X_2 and X_5 . The standard error of estimate (now designated as $\sigma_{1\ 2345}$ in full) equals 6.5, where before it was 6.7, a trifling change. The index of forecasting efficiency is now 28.4 per cent, where before it was 26.4 per cent. It is therefore questionable

whether the trouble of measuring and using in the regression equation the two additional variables is worth while. This is a good example of the way in which each additional variable yields diminishing returns in the way of improved predictions.

For the solution of the b coefficients, we introduce in Table 81 first the column headed σ_1/σ_k . This is the ratio by which each beta is to be multiplied. The b coefficients follow in column (6). They tell how many units X_1 is increasing for each unit of increase in the other variables. From these taken alone, it would seem that X_5 (interests) has the greatest bearing upon freshman marks and that X_4 (high-school average) has the least. But such is not the real situation. The best comparison of each variable's contribution to the variance in X_1 is to be seen in column (4), where each beta is multiplied into the corresponding raw r . Here it is seen that X_3 contributes nearly 22 per cent, X_4 nearly 17 per cent, whereas X_5 contributes only about 6 per cent, and X_2 about 5 per cent. These statements are relative to this correlational situation, with the influences of overlapping among the four taken into account. But as to choices among the four variables that we have here, they come in the rank order as per the βr products.

For the solution of the a coefficient, the last two columns are included. This coefficient turns out to be exactly 40.0. The entire regression equation now reads

$$X_1 = 40.0 + .182X_2 + .198X_3 + .142X_4 + .395X_5$$

With this equation, we could predict an X'_1 for every student, knowing his four scores in the other variables. As was said before, the addition of the terms involving X_2 and X_5 yield scarcely enough additional accuracy of prediction to justify their inclusion. One could try combinations of three predictive indices, variables X_2 , X_3 , and X_4 , or X_3 , X_4 , and X_5 , to see what happens. From the results in Table 81, it would seem that the last mentioned combination of three is the more promising. One could determine by another Doolittle solution whether it increased R sufficiently above .677 to justify its inclusion with X_3 and X_4 .

PARTIAL CORRELATION

The Meaning of Partial Correlation.—A partial correlation between two things is one that nullifies the effects of a third variable (or a number of other variables) upon both the variables being correlated. The correlation between height and weight of boys in a group where age is permitted to vary would be higher than the correlation between

height and weight for a group at constant age. The reason is obvious. Because boys are older, they are both heavier and taller. Age is a factor that enhances the strength of correspondence between height and weight. With age held constant, the correlation would still be positive and significant, because at any age taller boys tend to be heavier.

If we wanted to know the correlation between height and weight with the influences of age ruled out, we could, of course, keep samples separated and compute r at each age level. But the partial-correlation technique enables us to accomplish the same result without so fractionating data into homogeneous groups. When only one variable is held constant, we speak of a *first-order partial correlation*. The general formula is

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} \quad (100)$$

In a group of boys aged twelve to nineteen, the correlation between height and weight (r_{12}) was found to be .78. Between height and age, $r_{13} = .52$. Between weight and age, $r_{23} = .54$. The partial correlation is therefore

$$\begin{aligned} r_{12.3} &= \frac{.78 - (.52)(.54)}{\sqrt{(1 - .52^2)(1 - .54^2)}} \\ &= \frac{.4992}{.7190} \\ &= .69 \end{aligned}$$

With the influences of age upon both height and weight ruled out or nullified, then, the correlation between the two is .69.

As another example with three variables, the correlation between strength and height (r_{41}) in this same group was .58. The correlation between strength and weight (r_{42}) was .72. Although there is a significantly high correlation between strength and height, we wonder whether this is not due to the factor of weight-going-with-height rather than to height itself. So we hold weight constant and ask what the correlation would be then. Will boys of the same weight show any dependence of strength upon height? The correlation is given by

$$\begin{aligned} r_{41.2} &= \frac{.58 - (.72)(.78)}{\sqrt{(1 - .72^2)(1 - .78^2)}} \\ &= \frac{.0184}{.4343} \\ &= .042 \end{aligned}$$

By partialing out weight, it is found that the correlation between height and strength nearly vanishes. We conclude, therefore, that height *as such* has no bearing upon strength, but only by virtue of its association with weight does it show any correlation at all.

Second-order Partial.—When we hold two variables constant at the same time, we call the coefficient a *second-order partial r*. The general formula is

$$r_{12.34} = \frac{r_{12.3} - r_{14.3}r_{24.3}}{\sqrt{(1 - r_{14.3}^2)(1 - r_{24.3}^2)}} \quad (101)$$

In using this formula, as with others in this chapter, the subscripts will have to be modified to suit the choice of variables. Here we are assuming that we want to know the correlation that would occur between X_1 and X_2 with the effects of X_3 and X_4 eliminated from both. It is clear that this formula requires the solution of three partials of the first order previously.

As an example of this partial, we may cite the correlation between strength and age with height and weight held constant. This would mean that if a group of boys having the same height and weight were taken, would older boys be stronger? The raw correlation between age and strength was .29. The second-order partial also turned out to be .29. This means that it seemingly makes no difference whether we allow height and weight to vary or whether we do not; the relation between age and strength is the same within the range examined.

Some Suggestions Concerning Partial Correlation.—Needless to say, unless the assumptions necessary for computing the Pearson r 's involved are fulfilled, there is little excuse for using them as the basis for computing partial correlations. There are actually few occasions in psychology and education when a partial r is entirely defensible. The partialing out of such things as chronological age is perhaps the most common instance in which it is a useful device. It is not to be recommended as a lazy man's substitute for experimental control and fractionation of data. The newer processes of analysis of variance and tests of significance of statistics from small samples make experimental planning seem more important and the treatment of results more satisfactory without resort to partial correlations. It is inadvisable, in any case, to carry the partial-correlation method much beyond the first-order stage. Beyond this, the structure of the relationships becomes very much involved, and one is bringing more and more

raw r 's into consideration, each with its own fallibility. The building of an elaborate superstructure of statistics upon foundation stones that are not highly accurate in themselves can only lead to questionable results.

Exercises

1. Using Data *CC*, compute a regression equation involving X_1 , X_2 , and X_3 . Present beta coefficients, multiple R , and other necessary statistics. Interpret your results.
2. Using Data *DD*, compute a regression equation involving X_1 , X_3 , and X_5 . Present all statistics as computed in an ordinary solution to a multiple-correlation problem. Interpret your results.
3. Give Data *CC* a complete solution, using the Doolittle method. Include a regression equation and your interpretations.
4. Do the same for Data *DD* as was called for in Exercise 3.
5. Find the best combination of three predictive indices for either Data *CC* or Data *DD*.
6. For Data *CC* or Data *DD*, assume five reasonable sets of scores for five hypothetical individuals in the independent variables for which you have solved the regression equation, and from them predict X_1 values.
7. Determine the following partial r 's for Data *CC*: $r_{34.2}$, $r_{41.2}$, $r_{21.5}$, $r_{51.2}$. Interpret your results. Which of these coefficients have little meaning?
8. Determine the following partial r 's for Data *DD*: $r_{31.2}$, $r_{51.4}$, $r_{21.3}$, $r_{45.2}$, $r_{31.24}$. Interpret your results. Suggest other partial r 's that might be of importance to know about, and tell why.

DATA *CC*.—INTERCORRELATIONS OF SCORES FROM FOUR EXAMINATIONS AND MARKS RECEIVED IN FRESHMAN MATHEMATICS
($N = 100$)

Variable	X_2	X_3	X_4	X_5	X_1
X_2	..	.70	.53	.39	.51
X_3	.70	..	.61	.29	.51
X_4	.53	.61	.	.28	.61
X_5	.39	.29	.28	.	.39
X_1	.51	.51	.61	.39	.
M_x	4 10	5 44	5 37	4 95	5 70
σ_x	1 92	1 84	2 26	2 14	2 42

X_2 = Ohio State psychological examination.

X_3 = English-usage examination

X_4 = algebra examination.

X_5 = engineering aptitude examination

X_1 = marks in freshman mathematics

DATA DD.—INTERCORRELATIONS OF SCORES FROM FOUR EXAMINATIONS AND MARKS
 RECEIVED IN ENGINEERING DRAWING
 ($N = 154$)

Variable	X_2	X_3	X_4	X_5	X_1
X_2		53	24	28	33
X_3	53		24	11	34
X_4	.24	24		.38	.31
X_5	28	11	38		41
X_1	.33	34	.31	41	
M_x	4 19	5 42	4 70	4 85	5 25
σ_x	2 04	2 32	1 93	2 05	1 45

X_2 = Ohio State psychological examination

X_3 = algebra examination

X_4 = paper-folding test

X_5 = form-perception test

X_1 = term mark in engineering drawing.

CHAPTER XIV

RELIABILITY AND VALIDITY OF TESTS

The Importance of Reliability.—Much of what was said in previous chapters, when test scores were concerned, assumed that such measurements were perfectly reliable, or nearly so. By a perfectly reliable test, we mean one that is free from errors of measurement so that successive measurements of the same individual or phenomenon would yield exactly the same values. There are times, both in theoretical investigations and in practical work with tests, when it is very important to take into account the factor of reliability of scores. Conclusions to be derived from the same results might differ considerably whether or not we know the scores to be highly reliable or to be rather fallible. Many a conclusion in the literature is faulty because the apparent law of difference or lack of difference, as the case may be, may be entirely due to the unreliability of the measurements, which the investigator did not recognize or take into account. Thus the factor of reliability well merits special attention.

The Problem of Validity.—The question of validity of a test, or of test scores, is also a crucial one. The question has many facets, and it requires clear thinking not to be confused by them. In crudest terms, we say that a test is valid when it measures what it presumes to measure. This is one shade better than the definition that states that a test is valid if it measures the truth. In this chapter, it will be held that the problem of validity is a highly relative one. As a suggestive preview of the treatment to come, we may say that the question, "Is this test valid?" should be immediately answered by another question, "It is valid *for what?*"

RELIABILITY OF TEST SCORES

The Meaning of Reliability.—The main heading of this section, Reliability of Test Scores, puts the emphasis properly upon *scores*. It is the scores that are reliable or unreliable, as we shall see, not the test itself. For one reason, it is obvious that the same test could be scored in more than one way; in fact, many tests are so scored. Even when there is one established mode of scoring, the reliability coefficient obtained from a self-correlation of some kind will vary from one kind

of population to another. To speak of *the* reliability of a test is thus incorrect on these two counts. There are other reasons why reliability is a relative matter and why no absolute coefficient of reliability can be given for any test. Among the important reasons are the ways in which the reliability coefficient is derived. Three traditional ways of correlating a test with itself will now be described. They are the (1) test-retest method, (2) alternate-forms method, and (3) split-half method.

The Test-retest Method.—In repeating the same test in the same form with a group of testees, we encounter several factors that make this type of reliability coefficient of limited significance. If we could be sure that taking the test the first time left the testees just as they were before they took it, to face the second application of the test, we should be able to use this method with more assurance. With rare exceptions, what the testees learn during the first experience with the test is likely to carry over to the second trial. If the time interval is made long enough between applications of the test to take advantage of forgetting, intervening experiences become an important factor. Changes in the form of growth or of decline in individuals alter them with respect to the things measured. Experiments have shown that self-correlations of this kind undergo systematic reductions with the increase of time interval. If the meaning of this decline were known, we might be able to make certain corrections that would reduce all such self-correlations to the same quantitative basis. Unfortunately, the basic facts of change that affect the correlation here are not sufficiently known to justify such corrective attempts at present. Such self-correlations can be and often are obtained, however, as evidence of reliability, and they can be taken for what they are worth. In some instances, we actually may wish to know precisely this type of self-consistency of a set of test scores, particularly when we are interested in the stability of scores over a period of time. Then the retest method is the one to use.

The Alternate-forms Method.—If it is assumed that one can set up two or more comparable forms of a test—and this is often fairly well accomplished,—correlation of scores in the two yields an estimate of reliability. Here we face somewhat the same problems as in the retest method. Although the items are not identical in the two tests, the more comparable they are, the more opportunity there is for direct transfer of learning in the first-taken form to the second. Scores on the second-taken form are characteristically larger than those on the first. But if all testees profit by experience in proportion to their

original abilities at the time of the first test, or in equal degree, this systematic improvement need not interfere with a legitimate study of reliability. The question of time interval is again to be considered, and the results are much the same as with the retest method. Allowances must accordingly be made for this factor as well as others in interpreting the size of the obtained r .

The Split-half Method.—This method calls for the division of a test into two strictly comparable halves, two scores being obtained for every individual. The two halves constitute two alternate forms, but they are regarded as having been given simultaneously. As compared with the alternate-form method, the time interval between the two tests, is reduced to zero. The way to ensure most comparable halves, it is thought, is to constitute the one of the even-numbered items and the other of the odd-numbered items and to find an "odd-even" correlation, as it is called. Granting that there is no systematic alternation of items, so that odds and evens are genuinely random samples of items, this is probably as good a division as any. Within the two halves, the testee will have attempted an equal number of items or will have distributed his time rather evenly. In other divisions, such as first and last halves, this would not be the case.

The split-half method is generally accepted as the best of the traditional procedures, and it yields what we might call an "on-the-spot" estimate of reliability. It tells us of the accuracy of the scores at the time the individuals were measured, and from that we can infer that similar samples would be equally accurate under similar test conditions. The only flaw, which is readily corrected, is that the reliability coefficient varies with the length of test. A half test is not so reliable as a whole test under the same conditions. But by means of the Spearman-Brown formula, we can readily estimate what the reliability of the full-length test should be, if the two halves are really comparable. One important requirement in this instance is that the two parts should yield equal standard deviations in the same group of testees. The Spearman-Brown formula for estimating reliability when a test is doubled is

$$r_{11} = \frac{2r_{\frac{1}{2}\frac{1}{2}}}{(1 + r_{\frac{1}{2}\frac{1}{2}})} \quad (102)$$

where r_{11} = self-correlation of a test in its full length,

$r_{\frac{1}{2}\frac{1}{2}}$ = self-correlation of one-half of the test.

Rulon's Method.—Rulon has demonstrated that we may derive the reliability coefficient from split-half scores by using differences

between pairs of scores for individuals.¹ If the differences are all computed, squared, and then summed, by dividing by N , we obtain that part of the variance in the scores that is contributed by errors of measurement. Assuming, as is customary, that the total variance of the test scores can be divided into that part which is due to errors of measurement plus that which is attributable to true individual differences, we have a basis for estimating r_{11} through the relationship

$$r_{11} = 1 - \frac{\sigma_{1\infty}^2}{\sigma_1^2} \quad (103)$$

where $\sigma_{1\infty}$ = standard error of the obtained scores (see page 279)
and is calculated as described at the beginning of this paragraph

σ_1 = standard deviation of the total scores (half scores combined).

Rulon's formula is especially applicable when an IBM test-scoring machine is available, for this instrument can be so adjusted as to yield a difference between odds and evens for each testee. The method is subject to the same restrictions as any split-half procedure. It should be noted that *the formula gives the reliability of the total test scores and not of the halves*; so the Spearman-Brown formula need not be applied. If the Rulon difference formula should be applied to differences between scores on two forms, the reliability coefficient thus estimated applies to a test of twice the length of either form. A correction to the reliability wanted for each form can be made by substituting 5 for 4 in formula (111) on page 282.

Reliability Coefficients Based upon Rational Equivalence.—Considerable fault has been found with the split-half method, chiefly because each of the many rather equally acceptable ways of dividing a test into halves would yield different estimates of reliability. Which one of them, if any, should be regarded as the best estimate? In order to get around this objection and also the frequent difficulty of obtaining comparable halves, Richardson and Kuder² have recently derived new methods of estimating reliability. The reasoning behind these methods emphasizes the intercorrelations among the items themselves. No one wishes to compute all those intercorrelations for the sake of a single

¹ Rulon, P. J., A simplified procedure for determining the reliability of a test by split-halves, *Harv. educ. Rev.*, 1939, 9, 99-103.

² Richardson, M. W., and Kuder, G. F., The calculation of test reliability coefficients based upon the method of rational equivalence, *J. educ. Psychol.*, 1939, 30, 681-687.

reliability coefficient, but these authors have given formulas by which we can attain the equivalent result. Their most useful and yet accurate estimate of this kind of reliability is given by the equation

$$r_{11} = \frac{n}{n-1} \times \frac{\sigma_t^2 - \Sigma pq}{\sigma_t^2} \quad (104)$$

where r_{11} = reliability coefficient for the whole test.

n = number of items in the test

σ_t = standard deviation of the total test scores.

p = proportion of the group passing an item (or responding in some specified manner).

q = proportion failing to pass the item.

The product pq for each item is actually the *variance* of ability measured by that item, for the σ of an item is equal to \sqrt{pq} . The steps necessary for the solution of r_{11} in the Richardson-Kuder formula are as follows:

- Step 1. Determine the variance of the scores for the group of testees, in other words, σ_t^2 .
- Step 2. Determine for every item the proportion passing it (p) and the proportion failing it (q).
- Step 3. Determine the variance of each item (the product pq). Sum the pq products for all the items.
- Step 4. Substitute the known values in formula (104), and solve.

A Shorter Approximation to the Richardson-Kuder Reliability.—

If we are justified in assuming that all items have approximately the same degree of difficulty, which would be true when the proportions of individuals passing items are nearly the same, we may use a formula that is less demanding of information. It reads

$$r_{11} = \frac{n}{n-1} \times \frac{\sigma_t^2 - n\bar{p}\bar{q}}{\sigma_t^2} \quad (105)$$

where \bar{p} and \bar{q} with bars above them = average proportions of passing and failing testees to each item.

These can be obtained without counting successes and failures for every item, for the average p is equal to the mean of the total scores divided by n , and the average q is $1 - p$. From these facts, the formula can be simplified to

$$r_{11} = \frac{n\sigma_t^2 - \bar{R}\bar{W}}{(n-1)\sigma_t^2} \quad (106)$$

where \bar{R} = average number of right responses.

\bar{W} = average number of wrong responses (or $n - \bar{R}$).

\bar{R} is of course the mean of the total scores.

It should be said that both the formulas (104) and (105) slightly underestimate the reliability as compared with more exact ones, the second one more than the first. When we use them, we can be quite sure that the genuine reliability is at least as high as that which we obtained. It should also be said that these formulas are adapted to tests in which the testee receives one point for a correct response and none for a wrong response or an omission. When the scoring is weighted or when correction for chance success is introduced in scoring, other formulas may be used.¹

The Index of Reliability.—The self-correlation of a test is only one form of indicator of reliability. It indicates the closeness of relationship between scores from a test and other scores for the same individuals from the same test. The ability to predict an individual's score in the one application of the test from his known score in another application of the test depends upon the size of r_{11} . Another idea of reliability conceives of the "true" scores for a group of individuals and asks how closely the obtained scores correlate with these true scores. True scores in this instance are defined as those the testees would receive if the test were perfectly reliable. We can never know such true scores, yet because of certain statistical proofs, we can estimate the correlation between our obtained scores and the hypothetical true scores. This is possible because two comparable forms (or two halves) of the same test both are highly correlated with the true scores. And this correlation is definitely related to the correlation between the forms (or halves) by the equation

$$r_{1\infty} = \sqrt{r_{11}} \quad (107)$$

where $r_{1\infty}$ = correlation between obtained scores X_1 and true scores X_∞ .

The subscript for true scores is appropriately the infinity sign. The correlation $r_{1\infty}$ goes by the name *index of reliability* to distinguish it from the coefficient of reliability. It is the square root of the latter. If the self-correlation of a test is equal to .90, the index of correlation is $\sqrt{.90}$, or .95. The square root of a number between 0 and +1.0 is

¹ See Dressel, P. L., Some remarks on the Kuder-Richardson reliability coefficient. *Psychom.*, 1940, 5, 305-310. See also Hoyt, C. J., Note on a simplified method of computing test reliability. *Educ. and Psychol. Meas.*, 1941, 1, 93-95.

always larger than the number; so the index of reliability is always larger than the coefficient of reliability.

The Coefficient of Determination of Obtained Scores.—From equation (107), we find by squaring both sides that $r^2_{1\infty} = r_{11}$. It was previously stated (page 223) that the square of a coefficient of correlation between two things tells us the proportion of the variance in the one that is determined by the other. Since $r^2_{1\infty}$ is the coefficient of determination for a correlation between true scores and obtained scores, it tells us to what extent differences in the X_1 scores are determined by differences in the X_∞ scores, and so does r_{11} , to which $r^2_{1\infty}$ is equal. The coefficient of reliability therefore tells us immediately the proportion of *true variance* that there is in the obtained scores. The remaining proportion of the variance, which equals $1 - r_{11}$, is known as the *error variance*. If a test has an r_{11} equal to .90, we can say that 90 per cent of the variance in the obtained scores is true variance and 10 per cent is error variance. This is the only instance when a coefficient of correlation is treated as a percentage of something, and only then because r_{11} happens to be the square of another correlation, $r_{1\infty}$.

The Standard Error of an Obtained Score.—Since we can estimate the correlation between obtained and true scores and can think in terms of prediction of one from the other, we can also ask concerning the errors of prediction. We know the obtained scores and from them could predict true scores (assuming any mean and standard deviation we please for the true-score scale). But there is nothing to be gained by so doing, for the predictions would be no more accurate than the scores from which they were obtained, and nothing would have happened except a change of unit and zero point.

Suppose that we think in terms of prediction in the other direction; from true scores to obtained scores. This is impossible, since we do not know the true scores from which to make predictions. Let us think rather in terms of determination; of true scores *determining* obtained scores. But errors of measurement also help to determine obtained scores. We are interested in the extent of the discrepancies caused by these errors of measurement, in other words, in the size of distortions produced in the otherwise true-determined measurements. The average of these discrepancies is estimated by the formula

$$\sigma_{1\infty} = \sigma_1 \sqrt{1 - r_{11}} \quad (108)$$

where $\sigma_{1\infty}$ = standard error of an obtained score.

σ_1 = standard deviation of the distribution of obtained scores.

It will be seen that this is really the standard error of estimate of obtained scores from true scores. Because of the discussion just preceding, we preferred to talk in terms of determination of X_1 by X_∞ . This standard error therefore tells us concerning the amount of fluctuation to be expected in a score obtained for the same individual under very similar conditions. If, for example, a certain test has a standard error of an obtained score equal to 2.0 and a certain person has a score of 35, we could say, somewhat as we do about a mean when we know its standard error, that the odds are 21 to 1 that the truly determined score for this person would lie between 31 and 39 (plus and minus 2 SE 's from the obtained score). Such reasoning does not tell us what this individual's true score is, but it says how much distortion the errors of measurement may have injected into the obtained scores. The degree of confidence to be placed in individual differences in obtained scores is indicated to us by this standard error.¹

Reliability at Different Parts of the Test Scale.—Test users frequently ask to know the standard error of an obtained score rather than the reliability coefficient, because it tells them more directly what they wish to know. It tells them whether they should be concerned about differences of 2, 4, 8, or 12, points or whether any or all of these differences are within the probable range that could have been produced by errors of measurement. It may happen, however, that because of a peculiarity of the test itself, true discriminations are better at one part of the scale than at other parts. The σ_{1_∞} statistic is a blanket index, implying equal discriminating power all along the scale. If there is reason to suspect that true discrimination is unequal along the scale, this can be examined by preparing a scatter diagram, showing the relationship between two forms (or halves) of the same test. The standard deviations of the columns or rows at different score levels will indicate where predictions have the greatest accuracy (see page 193).

Computing the Standard Error of an Obtained Score from Differences. As was stated before (page 276), Rulon has pointed out the way of computing σ_{1_∞} directly from differences between scores made by individuals on odd and even pools of items. The equation is

$$\sigma_{1_\infty} = \sqrt{\frac{\sum d^2}{N}} \quad (109)$$

where d = any difference between two scores of half-tests for one individual.

¹ This statistic is frequently called the *standard error of measurement*

Reliability in Different Ranges of Measurements.—An examination of formula (103) on page 276 will show that if the standard error of an obtained score remains the same and the standard deviation of the tested group is increased, the ratio σ^2_1/σ^2_2 will become smaller, and so r_{11} will become larger. In other words, the larger the range of differences in the variable measured the higher the reliability. This is one important reason why we cannot properly speak of *the* reliability of a test, and it is also a reason why a knowledge of $\sigma_{1\infty}$ is often preferred to that of r_{11} . The standard error $\sigma_{1\infty}$ is relatively independent of the range of measurements, whereas r_{11} is not. If we wish to estimate the reliability coefficient in one range from the known reliability coefficient in another one of different scope, a formula similar to equation (89) on page 249 will serve:

$$\frac{\sigma_o}{\sigma_n} = \frac{\sqrt{1 - r_{nn}}}{\sqrt{1 - r_{oo}}} \quad (110)$$

where σ_o = standard deviation of the distribution for which the reliability coefficient is known.

σ_n = standard deviation of the distribution for which the reliability coefficient is unknown.

r_{oo} and r_{nn} = respective reliability coefficients.

If we know that a more limited group has a standard deviation of 8.0 and a reliability coefficient of .85 for a test, what will be the reliability coefficient in a more variable group whose σ is 10? Applying formula (110)

$$\frac{8}{10} = \frac{\sqrt{1 - r_{nn}}}{\sqrt{1 - .85}}$$

Squaring both sides of this equation, we have

$$.64 = \frac{1 - r_{nn}}{1 - .85}$$

Multiplying through by .15, we have

$$.096 = 1 - r_{nn}$$

and so

$$r_{nn} = .904$$

There is a standard-error formula for this estimated r_{nn} , but because standard errors for coefficients of correlation are of little or no value for such high r 's as reliability coefficients usually are and since one can then

almost always dispense with a test of the null hypothesis, such standard errors will not be presented here.

Reliability and the Length of Test.—It was indicated in connection with the split-half method that the whole test is more reliable than either half and that in general terms there is an increase in reliability going with increased length of test. This is true if the additional items added to a test are homogeneous with the ones to which they are added. By homogeneous, we mean that they have about the same intercorrelation with the items already in the test as those items have among themselves and possess about the same level of difficulty. If a test is increased A times, its present length, the reliability coefficient becomes

$$r_{AA} = \frac{Ar_{11}}{1 + (A - 1)r_{11}} \quad (111)$$

where A = ratio of the new length to the old.

r_{11} = reliability coefficient for the test of unit length.

To take a specific case, a test containing 50 items has a reliability of .80. What would be its reliability if we add 100 more items like the 50 we have? The solution is

$$\begin{aligned} r_{33} &= \frac{3(.80)}{1 + (3 - 1)(.80)} \\ &= \frac{2.40}{2.60} \\ &= .92 \end{aligned}$$

We can also predict reliability in a shorter test from the known reliability in a longer one. A then becomes a fraction, like $\frac{1}{2}$ or $\frac{1}{3}$. If we increase a test's length by 45 per cent, A is then 1.45. The formula will apply just the same as if we were dealing with simple ratios. If we knew the average intercorrelation among 40 items and wanted to forecast the reliability of the entire test, A would be 40.

Lengthening a Test to Attain a Certain Reliability.—We can use the Spearman-Brown formula in reverse. Knowing that the reliability of a short test is .75, we can ask how long the test would have to be to attain a reliability of .90. It is best first to solve equation (111) for A , which gives us the formula

$$A = \frac{r_{AA}(1 - r_{11})}{r_{11}(1 - r_{AA})} \quad (112)$$

Substituting the known values in this equation, we have

$$\begin{aligned} A &= \frac{.90(1 - .75)}{.75(1 - .90)} \\ &= \frac{.225}{.075} \\ &= 3.0 \end{aligned}$$

The test with $r_{11} = .75$ would have to be three times as long to attain a reliability of .90.

Any other level of reliability, larger or smaller, in which we are interested, can serve as our r_{AA} , and the necessary A ratio can then be computed. Experience will show that some tests of low reliability cannot reach some desired high reliability without being made indefinitely long, or so long as to be impracticable. Such a test would be given up without further extension and further work with it. Others will exhibit promising improvements in reliability with a moderate amount of extension. The formula is useful in this respect, that it helps decide upon rejection or extension of tests, or it is useful in cases in which a test is already too long for comfort whether or not shortening it would sacrifice too much reliability.

Reliability in Time-limit Tests.—Reliability indices of the split-half variety are most meaningful and dependable when derived from the kind of test or examination where every individual is allowed sufficient time to attempt every item. When r_{11} 's are determined for tests with a strict time limit, they should always be interpreted with caution. The odd-even division is a rarely justifiable mode of division into halves for a high-speed test. Let us suppose the extreme case in which the items are so easy that no one makes any mistakes, yet some finish 50 items and some only 20 in the time allowed because of rate of work. In this special instance, the coefficient would be 1.00—perfect reliability, if taken at its face value. It is not, of course, true that no errors are made in time-limit tests. But it is easy to see that to the extent that mere speed of work is important in determining the score, to that extent the apparent reliability is augmented. In other words, rigidly timed tests have an advantage over non-timed tests when compared for reliability.

There is no known correction for this, since we do not know for a specific test to what extent the speed factor has fostered a higher reliability coefficient. This is merely another instance of the relativity of reliability coefficients and should be kept in mind in comparing them. To a large extent, the same factor applies when the retest method and

alternate-form method are used. And it raises the more fundamental question of validity as to what abilities or traits are being emphasized in time-limit versus non-time-limit tests. To the extent to which errors determine individual differences, the time factor becomes less important. The time limit set upon a test, also, may have a bearing upon both its reliability and its validity. The issues are not yet very clear on these points, and experiments have not yet yielded all the answers to our problems in this area.

Reliability of Ratings and Other Judgments.—Many of the statistics described in connection with test scores also apply fairly well to human judgments of various kinds. The judgments may be in the form of rank order, rating-scale evaluations, paired-comparisons scaling, judgments in equal-appearing intervals, and the like. We can correlate the same observer's judgments obtained at two different times, or we can assume that similar judges are interchangeable and so intercorrelate their evaluations. We can pool judgments for two comparable groups of observers and correlate them so long as they apply to the same objects or persons. Experience has shown that with due cautions these applications may be made with meaningful results. Every coefficient must, as usual, be interpreted in the light of the manner in which it was obtained. Even the Spearman-Brown formula has been shown to apply, as, for example, in the pooling of judgments from two observers, which yields increased reliability in a manner found for the doubling of a test in length. The comparability of judges must be true here just as the comparability of items must be true in applying this formula to the change in length of test.

VALIDITY OF TEST SCORES

The Meaning of Validity.—Just as we cannot categorically answer the question, "How reliable is this test" neither can we answer the question, "How valid is this test" with any more singleness of meaning. There was a time, unfortunately still not entirely past, when each test was supposed to measure some underlying variable that went by a label. It was a test of intelligence, of introversion, or of neurotic tendency. Those concepts, because of the fixed labels, were supposed to be qualitatively fixed, known, and defined attributes. In order to be valid, tests going by those names were expected to correlate highly with older, generally accepted criteria of those supposed entities. For example, new tests were "validated" by demonstrating a strong correlation with the Stanford Revision of the Binet Test or with Laird's Test C2 or with Woodworth's Inventory

Now that these popular areas of personality have been shown to lack real unity and unanimity of reference,¹ we are properly more wary of attaching such labels to tests. If we regard intelligence as having been broken down into a collection of functional unities, called *primary abilities* for convenience, we find that the question of what is a valid intelligence test becomes meaningless. The primary abilities, on the other hand, have been arrived at by means of well-defined steps and can be verified by anyone who repeats those steps. If one acquiesces to the procedures by which those functional unities are discovered, he has no choice, if he still is concerned about the validity of tests, than to ask whether test *A* is a valid one for this primary ability or that one. The validity of a test as a measure of one of these factors is indicated by its correlation with the factor, which is its *factor loading*.² It is recognized by those who adopt the factor-analysis approach that no test is an unadulterated measure of any primary ability or trait. Not only is it diluted by errors of measurement, as we saw in the discussion of reliability but also it is contaminated to some extent with variances in other primary abilities or traits. This situation is overcome to some extent by combining tests that measure the same factor, with the idea that the minor impurities will tend to cancel each other in the process.

The vocational counselor and the vocational selector have faced a different kind of problem when they inquire about validity of tests. They are concerned about predicting outcomes in specified tasks and situations—clerical ability, scholastic ability, salesmanship, and the like. A test is a valid one for clerical aptitude if its scores correlate highly with later clerical proficiency. Another test is a valid one for aptitude in selling, because it correlates highly with later proficiency in selling. From this point of view, any test is valid for any sphere of behavior if it enables us to predict within that sphere, regardless of the name of the test or the supposed fundamental abilities that it measures. A test designed to predict the success of student aviators may prove also to be a valid test of scholastic aptitude in engineering or of aptitude for a military career in general. From the practical standpoint, the validity of a test is its forecasting efficiency in any measurable aspect of daily living.

¹ See in particular Thurstone, L. L., *Primary mental abilities Psychom. Monogr.*, 1938, 1, Guilford, J. P., and Guilford, R. B., *Personality factors D, R, T, and A. J. abnorm. (soc.) Psychol.*, 1939, 34, 21-36; and Mosier, C. I., *A factor analysis of certain neurotic tendencies. Psychom.*, 1937, 2, 263-286.

² For a brief discussion of factor theory and methods, see Guilford, J. P., *Psychometric methods*. New York: McGraw-Hill, 1936. Ch. XIV

Criteria of Validity.—One of the most difficult of all aspects of the validity problem is that of obtaining adequate criteria of what we are measuring. The factor-analysis approach has a fairly good solution when it is primary traits or abilities that we wish to measure. If two or more tests or items are combined to predict the factor, the validity coefficient is the multiple correlation between the tests and the factor. But practical criteria are most in demand and are most difficult to obtain and to measure adequately. An example of this is the criterion of scholastic achievement.

It has often been assumed that scholastic achievement, like intelligence, is a unitary attribute of each individual. But this is far from the truth. Although there is generally a positive correlation between achievement in different school subjects, there is sufficient disagreement to permit an individual to receive marks all the way from A to F in different subjects. It is best procedure, therefore, to examine the validity of each test used for guidance purposes in connection with *every* school subject taken by itself. Where a certain test of ability may possess only a moderate or low correlation with averages of school marks, it may correlate very high with specific courses. The writer has data showing correlations all the way from .37 to .74 between the Ohio State Psychological Examination, Form 20, and marks in freshman courses at a certain university. The point is that success in any sphere of life is ordinarily highly complex and is determined by many psychological factors in the individuals competing rather than one or a few. If we measure success in a complex activity by singling out as criteria one or more of its aspects and measuring them, we are checking upon the validity of the test or tests for predicting those chosen aspects. We should not identify those few aspects with the entire activity. We should, of course, attempt to single out the most significant aspects as criteria. Too often some inconsequential aspects are chosen because of their ready observability and measurability.

Having chosen the measurable variables of success in the area predicted, we have the problems of securing dependable measurements and perhaps of combining and weighting them in the wisest manner. With reference to measures of achievement, again, it should be emphasized that school marks as ordinarily assigned by teachers are rather poor metric material. Variations in meaning and standards from teacher to teacher and from course to course are notorious. Most marks are neither very reliable nor very valid indicators of achievement. The best measures of achievement in most courses are those obtained directly from good, comprehensive examinations of the objectively

scored type. Marks otherwise obtained often have reliabilities in the range from .60 to .80, and their validities are unknown. When we attempt to find the predictive value of a psychological test, therefore, shall we reject tests that fail to correlate highly with such fallible criteria? We can allow for the unreliability of criteria statistically when we know a coefficient of reliability for them. We cannot so easily know or allow for lack of *validity* of criteria, though we can make allowances, knowing the kind of criteria we have.

Correction for Attenuation.—When two fallible measures are correlated, the errors of measurement, if uncorrelated among themselves, always serve to lower the coefficient of correlation as compared with what it would have been had the two measures been perfectly reliable. We say that the degree of correlation has been attenuated. If we want to know what the correlation would have been if the two variables were perfectly measured, we must resort to the correction for attenuation, for which we have a formula

$$r_{\infty\infty} = \frac{r_{xy}}{\sqrt{r_{xx}r_{yy}}} \quad (113)$$

where $r_{\infty\infty}$ = correlation between two perfectly reliable tests (here tests X and Y , as indicated on the right-hand side of the equation).

r_{xx} and r_{yy} = reliability coefficients of the two tests.

The correlation obtained between a figure-classification test and a form-perception test was only .36. The reliability coefficients for the two tests were .60 and .94, respectively. Applying formula (113)

$$\begin{aligned} r_{\infty\infty} &= \frac{.36}{\sqrt{(.60)(.94)}} \\ &= \frac{.36}{.751} \\ &= .48 \end{aligned}$$

We should therefore expect the correlation between true scores in these two tests to be .48 rather than the obtained one of .36. In general, when making this correction for attenuation in both fallible tests, if we are dealing with two forms of the same test for purposes of finding reliability, there is a possibility of determining four intercorrelations between the two tests; *i.e.*, each form of the one correlated with the two forms of the other. In this case, it is well to use all the information we can get concerning the intercorrelation of the two tests by computing

the four coefficients and using their arithmetic mean as a better estimate of the numerator of the fraction in formula (113).

Correction for Attenuation in the Criterion Only.—The preceding device has limited application except in theoretical problems. In practice, we are compelled to deal with fallible tests. If the tests from which we wish to predict something else are not perfect, that fact must be faced, and our predictions are reduced in accuracy accordingly. But we should hardly expect to be asked to overlook the fallibility of the criterion we are trying to predict. If it measures success inaccurately, this lack of accuracy should not be permitted to make it appear that the test is less valid than it really is. It is becoming more customary, therefore, to correct validity coefficients for attenuation in the criterion measurements but not in the test scores. This one-sided correction is made by the formula

$$r_{\infty x} = \frac{r_{xy}}{\sqrt{r_{yy}}} \quad (114)$$

where $r_{\infty x}$ = correlation between X and Y with errors of measurements allowed for in the criterion Y but not in the test X .

As an application of this formula, we cite a line-drawing test that correlated with a teacher's rank-order judgments of creative ability in her students in design to the extent of .65.¹ The reliability of the teacher's ratings (combined from two rank orders a month apart) was found to be .82. Had the teacher's ratings been perfectly reliable measures of the thing she was judging, the correlation with test scores would have been $.65/\sqrt{.82} = .72$. The correlation of .72 is accordingly taken as the genuine validity of the test, unless we are concerned about predicting teacher's judgments, contaminated by flaws as they obviously are, rather than genuine ability as evidenced by those ratings.

Standard Error of the Estimate of a True Criterion.—Taking the correlation between our fallible scores and an infallible or true criterion as the coefficient of validity, we shall also have smaller errors of prediction than if we tried to predict fallible criterion measurements. We could substitute $r_{\infty x}$ in the usual formula for finding the standard error of the estimate from r , but the σ_{yx} (which now becomes $\sigma_{\infty x}$) can be calculated directly from the original correlations by the formula

$$\sigma_{\infty x} = \sigma_y \sqrt{r_{yy} - r_{yx}^2} \quad (115)$$

¹ Guilford, J. P., and Guilford, R. B. A prognostic test for students in design. *J. appl. Psychol.*, 1931, 15, 335-345.

The Index of Forecasting Efficiency with a True Criterion.—An index of forecasting efficiency could also be computed directly from $r_{\infty x}$ to denote the improvement in predicting the true criterion variable on the basis of knowledge of test scores over prediction without that knowledge (see page 223). This statistic can be calculated directly from the known r 's, however, without first finding $r_{\infty x}$, by use of the formula¹

$$E_{\infty x} = 100 \left(1 - \sqrt{1 - \frac{r_{yx}^2}{r_{yy}}} \right) \quad (116)$$

Validity and Length of Test.—Since the more reliable a test becomes (less errors of attenuation) the higher its validity coefficient will be, it follows that, everything else being equal, the longer a test the greater its validity. If change in length is equal to some ratio A (length of new test divided by length of old test), the new validity coefficient is given by the formula

$$r_{y(Ax)} = \frac{r_{yx}}{\sqrt{\frac{1 - r_{xx}}{A} + r_{xx}}} \quad (117)$$

where $r_{y(Ax)}$ = validity coefficient between criterion Y and test X , now A times as long as originally.

r_{yx} = uncorrected validity coefficient.

r_{xx} = reliability coefficient for test X .

For the sake of an illustration, suppose that the line-drawing test already referred to, with $r_{yx} = .65$ and $r_{xx} = .57$, were to be made twice as long by adding comparable items. The new validity would become

$$\begin{aligned} & \frac{.65}{\sqrt{\frac{1 - .57}{2} + .57}} \\ &= \frac{.65}{.886} \\ &= .73 \end{aligned}$$

It would thus definitely pay to make this test longer. If we were considering validity in predicting the true criterion, by formula (114), the estimated validity coefficient would become .81.

¹ Conrad, H. S., and Martin, G. B. The index of forecasting efficiency for the case of a "true" criterion. *J. exp. Educ.*, 1936, 4, 231-244.

Shrinkage of Validity in New Samples.—Lest the recent comments lead to too rosy a picture of validity of tests, we must now consider factors that encourage less optimism. In recent paragraphs, we have been concerned about the fallibility of criteria and about making allowances that would give tests their due credit. All that has been said about validity assumes that the sample from which our statistics were obtained is a representative random selection of the population within which the test or tests will be used for predictive purposes. Even if this were true, sample statistics have the habit of overestimating to some extent the estimates of dependability of predictions even within the population sampled, particularly when N is small. What is worse, tests are often applied to new populations, of which our sample is not a very good representative. The extent to which similar dependability of predictions can be obtained will hinge upon the degree of similarity of sample and new population. The wise counselor or engineer will make all due allowances for these risks and will, if possible, recheck for validity of the tests he uses in new connections. Unfortunately, no fixed rules can be laid down at this time to indicate the amount of shrinkage to be expected under varying circumstances of test application.

The Validity of Test Batteries.—When diverse tests are combined in a battery to predict some practical criterion, the question of weighting the parts comes into the picture. One general solution has been to derive a multiple-regression equation and to weight each test according to its b coefficient. A multiple coefficient of correlation then gives us an index of validity and a standard error of estimate, an index of accuracy of prediction of the criterion. If this is the solution, certain general rules can be offered to aid in the selection of tests that together will yield the greatest validity coefficient.

One rule is that each test should correlate high with the criterion and low with the others. A common-sense explanation of this is that tests that correlate high with each other thereby duplicate one another in predicting the criterion. The practical criterion is usually a composite affair, made up of variances in different abilities and traits. To obtain maximum coverage of this area, the parts of a battery should not overlap any more than is necessary.

But this is not the whole story. Sometimes a part of a battery can correlate low, even zero with the criterion, *but has predictive value in a battery if it correlates very high with another test that does correlate with the criterion.* In a multiple-regression equation, such a test will probably end with a negative b coefficient, but any weight, negative

as well as positive, that deviates significantly from zero contributes something toward prediction. What probably happens is that the negative weight serves to cancel some "foreign" variance in the test that correlates with the criterion—"foreign" in the sense that it is that part of the test which is not in common with the criterion

The Problem of Weighting Parts in a Battery.—Multiple-regression weights may not always be the best ones for the parts of a battery. Simply summing scores without weighting obtained scores sometimes yields as good results as with the use of b coefficients. Certainly most b coefficients are awkward to use in practice, because they are not simple integers. If weighting is to be based upon this principle, small integral values will serve just as well. For example, if the three b coefficients in a certain battery are 1.12, 2.35, and .65, we might call the smallest one a weight of 1, and the others will be simple multiples of the smallest. They would become 2 and 3 for 1.12 and 2.35, respectively. The effect upon the multiple correlation would probably be trivial, and the arithmetic of prediction would be very much simplified.

Even when part scores are merely summed, it must be remembered that the tests are weighted. The variance of the total score is made up of the variances of the parts, which themselves are unequal. Other things being equal, the greater the variance of the part, the greater its contribution to the total. To allow for this, some testers have resorted to weighting each part inversely as the standard deviation of that part. The effect is as if every part score were reduced to standard-score form before it is combined with others. This practice is justifiable only when the parts have approximately equal reliability. The reason is that by this procedure the test with the smallest standard deviation receives the largest weight, and the smallest standard deviation may be associated with unreliability, because the test is a short one. In the absence of knowledge and experience that would support a particular kind of weighting of parts, the investigator would do well to play safe by merely summing point scores, which practice has also to its credit the advantage of simplicity.¹

When Test Scores Should Not Be Combined.—Sometimes it is best not to consider pooling parts of a battery to predict a single criterion at all. If the parts are really diverse variables intended to cover a complex area of life success, weakness in any one of them might

¹ For a very recent advanced discussion of the general problems of combining test (and items) see Richardson, M. W., *The combination of measures, in The prediction of personal adjustment*. New York: Social Science Research Council, 1941. Pp. 379-401.

be serious. And yet in a combined score where all other parts are high, this weakness would be glossed over. Another approach would therefore be to set up critical score limits for all parts taken separately. Any applicant or counselee falling below the critical score in any one part would then be in the doubtful category. If a limited number of applicants are to be selected out of a great many, as in civil-service procedures, high critical scores could be determined for each test variable, and anyone falling below in any variable would be eliminated. For each test, critical division points could be set up after the manner of predicting attributes from scores (see pages 181*ff.*), provided that criterion subjects could be classified as being in one or more of a few categories. Where selection indicates consistently high performance of several kinds or where a defect of any kind cannot be compensated for by higher status in other qualities, this general type of prediction is called for. When there are several criterion variables instead of one composite one, then the consideration of part tests by themselves is all the more in order. The use of profiles is also recommended when test scores are relatively uncorrelated.

ITEM ANALYSIS

Many of the comments just made concerning the combining of tests into batteries also apply to the problems connected with combining items into a test. But the examination of a test, item by item, brings up a host of additional problems to which we shall now give our attention. The subject is entirely too extensive for us to do more than introduce it here and to give some of the answers to the more pressing and universal questions that arise. The subject divides itself into two general but not unrelated problems: (1) the problem of item difficulty and (2) the problem of item validity.

The Difficulty of Test Items.—The customary procedures for evaluating level of difficulty of test items on a linear, rational scale were described briefly in an earlier chapter (see pages 114*ff.*). The principle underlying the methods is that the proportion of a group passing an item marks off a division point in the curve of normal distribution. From that proportion, which represents the area under the curve lying above the point of difficulty on the base line, a corresponding deviate value can be readily determined. In connection with these procedures, it is necessary only to repeat for emphasis here the fact that difficulty values are always relative to the distribution of ability in the sample tested and also the fact that when items have

only a few alternative answers, a correction for chance success must be made (see page 116).

A knowledge of difficulty of items serves several purposes. One of long-standing recognition is the practice of arranging items in rank order in a test, easiest items first. If there is felt a need of graduating the difficulty more or less steeply from beginning to end, with a liberal supply of items from which to draw, this need can be fulfilled. It is also important to temper the general average difficulty of items to the ability level of the group to be examined. Here several accepted rules apply, rules that are wise on both theoretical and empirical grounds. Items passed by everybody or failed by everybody are of no value for measurement purposes. This rule may be violated for the sake of introducing one or two very easy "shock absorbers" at the beginning of a test. The maximum discrimination among testees is to be obtained by items that about one half the individuals can pass. This rule implies proportions that have been corrected for chance success. In true-false tests, this would mean items passed by 75 per cent of the individuals. Because tests of this average level of difficulty are sometimes discouraging to testees, for the purpose of maintaining better morale, the rule may have to be violated somewhat by lowering the general level of difficulty.

Level of Difficulty and Validity of a Test.—The validity of a test may be seriously influenced by its level of difficulty. In subjecting the Seashore Test of Pitch Discrimination to a factor analysis, the writer discovered that at different levels of difficulty three distinctly different abilities were being measured.¹ It is probable that among the easiest items, where differences between pairs of tones were 17, 23, and 30 cycles per second, individual differences in scores represented differences in attentiveness to an easy task, and errors were made because of lapses of attention. Some of the more difficult items, having differences of 1 to 5 cycles, were most heavily correlated with a second factor or ability. Items with differences from 3 to 12 cycles were most heavily correlated with a third factor or ability. The moral is that a total score limited to one of the three ranges would be most valid for that particular range. A total score based upon the total range of differences from 1 to 30 cycles would represent some kind of composite measurement.

It is likely that other tests, likewise, are altered in validity as they are easy or difficult for the group examined. Very easy tests may

¹ Guilford, J. P., The difficulty of a test and its factor composition. *Psychom.*, 1941 6, 67-77.

become measures of perceptual ability or of motor speed, and more difficult tests of the same kind of items may be measures of reasoning power of one kind or another. More attention must therefore be paid in the future to the appropriate level of difficulty of a test for the group to which it is administered when we want to be sure of its validity.

The Diagnostic Value of Items.—The heart of the item-analysis problem is the matter of diagnostic value of items. To be diagnostic of any trait, an item must enable us to distinguish between individuals who have more or less of that trait. Those who respond in one way to the item must, on the whole, be different in the trait from those who respond in other ways. In the case of abilities, those who pass must have significantly higher ability of the kind we wish to measure than those who fail. We therefore have a prediction problem. Some items cut with more precision among different kinds of people; others show up little or no difference among them. As with total tests, we wish an item to predict some criterion.

The Criterion of Internal Consistency.—In testing the validity of items, we recognize two kinds of criteria. One is the usual kind of outside criterion, which we also correlate with total test scores. The other consists of provisional test scores, usually total scores in the test of which the item is itself a part. We can set up a test that, because of prior knowledge, we believe can be used to measure some trait. We can score each item in accordance with preconceived ideas of what kinds of response indicate more of the trait. Taking the total score as our provisional criterion, we can correlate each item with the criterion in ways soon to be described. Every item that correlates significantly with this criterion may be adopted as being diagnostic. We are in this way bringing to the items the well-known test of internal consistency.

A number of cautions and restrictions should be mentioned in connection with this process. In the first place, we should be very certain as to what the criterion scores represent. It has too often been assumed that because in the end we have only items that do possess considerable internal consistency or correlation with the same criterion, they are diagnostic of some unitary trait. This is not necessarily correct. The total collection of items, although centering about some hypothetical unitary trait like introversion-extraversion, is usually a measure simultaneously of *several* real variables in personality. Unless a factor analysis or something equivalent has been made to establish the unity of the trait, the trait itself is probably complex. Even if it is shown that the items correlate with the total, there may be sub-clusters that correlate among themselves but not so much with items in other clusters.

One item correlates with the total score because it has in common with the total score validity for trait P , but another so correlates because of its correlation with trait Q . Nor does the fact that the items do not correlate with test scores outside this one or the fact that total scores do not correlate with one another point to the fact that each test measures a single variable. The route to the measurement of unitary variables cannot be taken by way of tests of internal consistency.¹

Another minor difficulty is that the item itself helps to determine the total score, and we are dealing with correlation of part with whole. When the test is a long one, more than 50 items, this is of trivial consequence. It probably calls for some kind of correction or allowance when the number of items is 20 or less.

The size of sample from which the item validities are determined must be rather large if we want stable results that can be generalized. If the criterion group is divided into two subgroups, upper and lower, there should be about 100 individuals in each group, whether the subgroups represent upper and lower quarters or halves of the total distribution of criterion scores. Smaller samples will do to indicate tendencies, but fluctuation of correlations between items and criterion from sample to sample is something to be seriously considered.

Indices of Item Validity.—By far the most common index of validity for a test item, whether the criterion is for the purpose of testing internal consistency or whether it is some external criterion, is some type of correlation coefficient. The most accurate of these, and also one of the most laborious, is the biserial r (see page 237). In this case the criterion group is divided into two subgroups, those passing and those failing to pass the item, and on the assumption of normal distribution of ability to pass the item, a correlation with total criterion scores is computed. Since the particular individuals and the number of them who pass will differ from item to item, it will be necessary to sort them out for every item, to form a distribution, and to compute a mean and standard deviation. The mean and standard deviation of the total group will serve alike for all items. In arriving at a critical point or lowest acceptable biserial r , it is well to make the null hypothesis, compute the standard error of the biserial r when $r_b = 0$. An item that correlates with the criterion more than two standard errors may be regarded as having significant validity, whereas one correlating more than 2.6 SE 's has very significant validity, to use Fisher's fiducial limits.

¹ Sletto, R. F., *Construction of Personality Scales by the Criterion of Internal Consistency*. Minneapolis: Sociological Press, 1937.

Two other types of correlation coefficient used for the same purpose are the tetrachoric r (see page 240) and the phi coefficient (see page 245). The tetrachoric r requires a very large number of cases, preferably more than 300, in order to give reasonably stable coefficients, and without facilitating tables the work of computation is almost prohibitive.¹ The phi coefficient, although not always giving values equivalent to Pearson r 's, has some practical advantages. Indeed, when the two criterion subgroups are equal in number, if we know the proportion of each group passing the item (or reacting to it in some specified manner), the formula for ϕ simplifies to,²

$$\phi = \frac{p_u - p_l}{2\sqrt{pq}} \quad (118)$$

where p_u = proportion of the upper criterion group that responds in some specified manner to the item.

p_l = proportion of the lower criterion group that responds in the same manner.

p = proportion of the two subgroups combined that react in this manner and is given by the relation $p = \frac{p_u + p_l}{2}$.

$q = 1 - p$.

An *abac* for the solution of the phi coefficient is given in Fig. 41. All one needs to know to use this *abac* are p_u and p_l . Looking for the ordinate and abscissa values corresponding to p_u and p_l for an item, we find that the intersection of these lines will locate ϕ , which can be estimated to the second decimal place between curved diagonals. The chief objection to ϕ as a measure of correlation is that it is not always equivalent to the Pearson r . Since in any item-analysis problem, it is the relative validity coefficients that we want, this is not an important objection. Another objection is that the size of ϕ will vary according to whether we include in the subgroups 50 or 25 per cent, or any other percentage, of the cases. So long as we stay with the same percentage in upper and lower subgroups, the ϕ coefficients will be comparable. It has been the writer's experience, without

¹ See Flanagan, J. C., General considerations in the selection of test items, etc. *J. educ. Psychol.*, 1939, **30**, 674-680, where an *abac* is given for the graphic solution of Pearson r 's when the criterion subgroups are the highest and lowest 27 per cent of the distribution. See also Mosier, C. I., and McQuitty, J. V. Methods of item validation and abacs for item-test correlation, etc. *Psychom.*, 1940, **5**, 57-65.

² See Guilford, J. P., The phi coefficient and chi square as indices of item validity. *Psychom.*, 1941, **6**, 11-19.

mathematical proof for it, that when highest and lowest quarters are used, ϕ is equivalent to the Pearson r . And when highest and lowest halves are used, a Pearson r may be estimated, if desired, by means of the methods suggested previously (see page 253). The test of the null hypotheses and establishment of the lowest significant ϕ 's can be

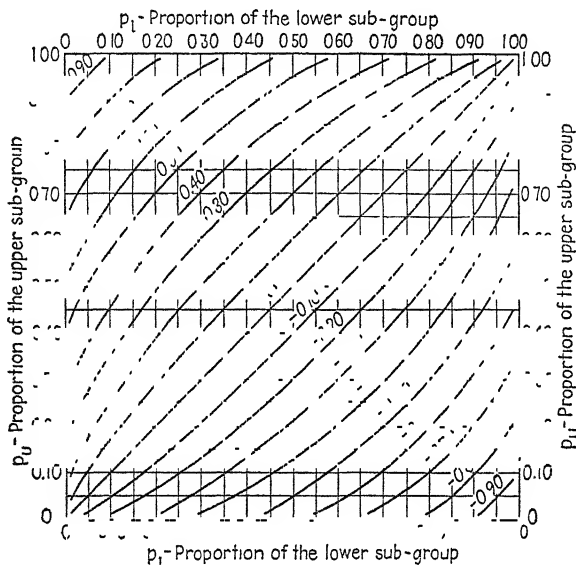


FIG. 41 —An abac for graphic estimates of the phi coefficient when one variable has an even division of cases in two categories. If the proportion of the upper criterion group passing an item is 0.65 and the proportion of the lower group passing it is 0.30, ϕ is found at the intersection of the horizontal line at level 0.65 and vertical line at level 0.30. It is midway between the lines for $\phi = 0.30$ and $\phi = 0.40$, so the value we are looking for is 0.35.

accomplished through the use of chi square. When the two criterion subgroups are equal in size

$$\chi^2 = N\phi^2 = \frac{N(p_u - p_l)^2}{4pq} \quad (119)$$

where the symbols mean the same as in formula (118). For the case of 1 degree of freedom, which we have in a fourfold table, a chi square of 3.841 is considered significant, and one of 6.635 is very significant. A significant phi coefficient would therefore be equal to

$$\sqrt{\frac{3.841}{N}} \quad (120)$$

and a very significant phi is equal to

$$\sqrt{\frac{6.635}{N}} \quad (121)$$

From what has just been said, it is seen that chi square is another indicator of validity of test items. It carries with it its own test of significance, but in cases of fourfold tables of item-test correlation, it is not so convenient to compute as ϕ . When there are three or more categories of responses for an item rather than two, the chi-square test would also apply.

Another index of validity sometimes used is the critical ratio. The difference between means of passing and failing subgroups (or of two groups otherwise distinguished because they have reacted differently to an item) is divided by the standard error of the difference. The difference between proportions of two criterion subgroups (highest and lowest quarters or halves) may also be compared to its standard error. Critical ratios of 2.0 and sometimes 3.0 are demanded of an item that is to be retained. Mosier and McQuitty have prepared an abac¹ from which the ratio may be read when we know the two proportions. The size of critical ratio depends upon N , the number of cases in the subgroups, as well as upon the difference between means or proportions. Accordingly, critical ratios are directly comparable for size only when the number of cases in the criterion groups is constant. An advantage in its use is that, regardless of the number of cases in the sample, it carries its own indication as to significant deviation from the null hypothesis.

Weighting Responses to Test Items.—Many of the problems that apply to the weighting of tests in a battery also apply to the weighting of items in a test. If we followed completely the principle of the multiple-regression equation, we should have to correlate the items with each other as well as with the criterion. The usual number of items in a test makes this procedure prohibitive. In predicting a practical criterion, the correlation of each item with the criterion should be as high as possible and the intercorrelations as low as possible, or else an item that correlates zero with the criterion must have a high correlation with some other item or items in order to be of value in predicting the criterion.

Procedures that approximate the principles of the regression equation have been suggested from several sources. The most efficient

¹ Mosier and McQuitty, *op cit.*

method of this kind seems to be that described by Richardson and Adkins.¹ The method yields a weight for each item by the application of the formula

$$L = \frac{r_{yi} - r_{xi}r_{xy}}{(r_{xy} - r_{yi}r_{xi}) \sqrt{pq}} \quad (122)$$

where L = weight (so called because it is an approximation to the weight computed by Toops' longer L -method).

r_{yi} = correlation between the item and the criterion

r_{xi} = correlation between the item and the total test score in which this item is a part.

r_{xy} = correlation between total test score and criterion.

p = proportion that pass the item (or react to it in some specified manner).

$q = 1 - p$.

If we disregard the intercorrelations among items, or assume that they are approximately equal, we may apply a scoring weight that was introduced by the author. It is proportional to the correlation between item and criterion (r_{yi}) and inversely proportional to the variance of the item (pq). The formula is²

$$W = \frac{p_u - p_l}{pq} + 4 \quad (123)$$

where the symbols are defined as in formula (118). This formula yields weights ranging from 0 to 8, with a weight of 4 when the item-criterion correlation is zero. The formula applies particularly in the case where the two subgroups are equal in numbers. When they do not happen to be equal, to find p_u and p_l for the two groups respectively has the effect of equalizing their importance. A standard error has been provided for the special case when the item-criterion correlation equals zero (or when $W = 4$). This is the case of the null hypothesis. The formula is

$$\sigma_w = \frac{2}{\sqrt{Npq}} \quad (124)$$

where N = number of cases in the two subgroups combined.

p and q are as defined in the preceding equation.

¹ Richardson, M. W., and Adkins, D. C. A rapid method of selecting test items *J. educ Psychol.*, 1938, **29**, 547-552.

² Guilford, J. P. A simple scoring weight for test items and its reliability. *Psychom.*, 1941, **6**, 367-374.

Let us apply the last two formulas to a particular item. It is the question, "Would you rate yourself as an impulsive individual?" from a personality inventory that attempted to score individuals for degree of depression. From provisional scoring, the highest and lowest quarters of a group of 1,000 students had been segregated, the former being designated as the "depressed" subgroup and the latter as the "not-depressed" subgroup. Table 82 shows the proportions of the two subgroups responding by saying "Yes," "?" and "No" to the question. The work of solving for the weight to be assigned to

TABLE 82 — THE SOLUTION OF SCORING WEIGHTS FOR RESPONSES TO AN INVENTORY QUESTION
($N = 500$)

	Responses			
	Yes	?	No	
p_u	.284	180	.532	Depressed Subgroup
p_l	.424	140	.436	Not-depressed Subgroup
p	.356	160	.484	Both Combined
$p_u - p_l$	- 140	+ 040	+ 096	
pq	2293	.1344	2497	
\bar{W}	3.4	4.3	4.4	
σ_W	.19	24	18	

each response is briefly outlined in the lower rows of Table 82. The three weights are 3.4, 4.3, and 4.4, for the three responses, in the order given. If we use only integral weights we should have to round them to 3, 4, and 4, respectively. Only the response "Yes" deviates far enough from 4 to be rounded to anything but 4.

The standard errors for deviations from 4 (null hypothesis) for the various responses are given in the bottom row of Table 82. Two deviations are statistically significant, but in view of the fact that only one of the responses had a weight deviating as much as barely more than half a unit from 4, it would probably be of little value to keep this item to help predict depression. It may be diagnostic of some other trait that the same weighting formula would reveal. For the

student's convenience, an abac for estimating integral weights on the basis of formula (123) is presented in Fig. 42. Its chief usefulness is to be found in connection with weighting responses to personality-test items, but it also applies to items intended to assess abilities.

The Importance of Weighting Item Responses.—Personality tests in the past have generally leaned heavily upon a weighting system of

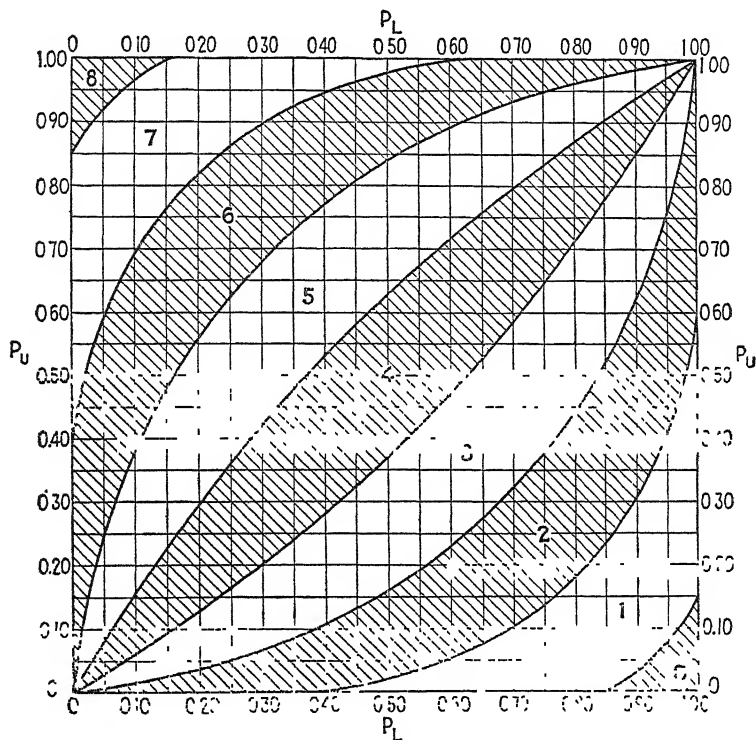


FIG. 42.—An abac for the graphic determinations of scoring weights for responses to test items. If 65 per cent of an upper criterion group responded in a certain manner to an item and 25 per cent of a lower group responded in the same manner, the intersection of the two lines (0.65 on the vertical scale and 0.25 on the horizontal one) falls within the band where $W = 6$. The scoring weight for this response to the item is 6.

some kind. Typical of this statement are the Strong Vocational Interest Test and the Bernreuter Personality Test. There are instances in which weighted scoring has materially improved reliability over that attainable with unweighted scoring. By "unweighted scoring," here we mean that responses are given a value of 0 or 1 only. Tests of validity have generally not shown so much benefit from differential weighting of items. Every test constructor, in these days of machine

scoring, in which differential weighting is bothersome, should be challenged to show good cause for other than the simplest system of weighting. It is more important to be sure that items are significantly correlated with criteria and to reject those not significantly correlated. And in order to make an adequate examination of significance, we are called upon to use large samples that will serve us with reliable tests of item significance. In this way lie the selection of dependable items and the construction of dependable tests that can be safely and meaningfully applied to individuals beyond our validating sample.

Exercises

- 1 The following reliability coefficients were presented for a certain test

Split half	96	Retest after 1 month	91
Alternate form	94	Retest after 2 years...	86

Are these coefficients reasonable? Discuss.

- 2 In six tests, the following correlations were found between halves composed of comparable items: .43, .55, .66, .74, .86, and .94.

- Determine the reliability coefficient for the full-length tests
- Determine the index of reliability in each case

- 3 In a certain test, the sum of the squared differences between scores on two comparable halves equaled 285. $N = 50$, and $\sigma = 8.5$. Find the coefficient of reliability for the total scores and the standard error of estimate of the obtained scores.

- 4 In a test of 55 items, the standard deviation of the total scores was 7.5. The sum of the variances of the items was 9.8327. Find the reliability coefficient.

- 5 Another test of 150 items has a standard deviation of total scores equal to 24.4 and a mean of 94.2. Find the reliability coefficient, assuming that the items are approximately equal in difficulty.

- 6 In four tests, the reliability coefficients were .65, .76, .87, and .94. Determine $r_{1\infty}$ and $\sigma_{1\infty}$ in each case, assuming that $\sigma_x = 10.0$.

- 7 A test has $\sigma_x = 7.2$ and $r_{11} = .86$. In another group, σ_x is 6.0. What will the reliability become? In still another group, $\sigma_x = 9.0$. What will r_{11} become?

8. Complete the following table by performing the necessary computations.

r_{11} \ A	1.5	2	4	6	10	20
30	.39					.90
70			.90	.93		
.90	.93	.95			.99	

9. For the data in the last exercise, plot on graph paper the increase in r_{AA} (on the ordinate) as A increases (on the abscissa) for each value of r_{11} . Draw some general conclusions from the table or from the diagram

10 Complete the following table by computing the necessary A 's.

$r_{11} \backslash r_{AA}$.65	.75	.85	.95
30				
50				
70				
90				

11 Test X has a reliability coefficient of .92, and criterion Y (a final examination of the essay type) has a reliability of .65. Assuming that the validity coefficient in four trials had values of .35 and .48, .61 and .72,

- Determine the probable correlations between "true" measurements in test and criterion
- Determine the genuine validity of the fallible test in each case (assuming a perfect criterion)

12. In the preceding problem, assume that $\sigma_y = 15.0$. Compute $\sigma_{\infty x}$ and $E_{\infty x}$ for the four instances.

13. Four tests have reliability coefficients and validity coefficients as follows.

Test	X_1	X_2	X_3	X_4
r_{xx}	.80	.80	.60	.80
r_{yx}	.70	.50	.50	.30

Determine the validity in each case, assuming that each test is doubled in length. Do the same, assuming that each test is made five times as long

14. Determine for the items in Data EE one or more of the indices of validity mentioned in this chapter. Tell something about the significance of each index, and draw any other conclusions that suggest themselves.

DATA *EE*—FOR 10 TEST ITEMS ARE GIVEN THE PROPORTIONS OF THE INDIVIDUALS IN UPPER AND LOWER QUARTERS OF A CRITERION GROUP WHO PASSED EACH ITEM AND THE NUMBER OF CASES IN THE TWO SUBGROUPS COMBINED (N)

Item	p_u	p_l	N
1	84	64	50
2	90	56	50
3	80	.45	100
4	75	58	100
5	.15	21	200
6	.45	27	200
7	.62	52	400
8	.96	90	400
9	47	40	1000
10	56	61	1000

15. Give evidence of the degree of validity of the items in Data *FF*. Are all responses equally diagnostic? Discuss

DATA *FF*—PROPORTIONS OF TWO CRITERION SUBGROUPS WHO RESPOND IN ONE OF THREE WAYS TO TWO QUESTIONNAIRE ITEMS
($N = 500$)

Question 1. Do you daydream frequently?

Group	Yes	?	No
Low cycloid	46	09	45
High cycloid	.71	07	22

Question 2. Do you consider yourself less emotional than the average person, *i.e.*, less easily upset?

Group	Yes	?	No
Low cycloid	55	.04	.41
High cycloid30	04	66

16. Determine scoring weights for the three responses to the two items in Data *FF*.

APPENDIX

TABLE A—SQUARES AND SQUARE ROOTS OF NUMBERS FROM 1 TO 1,000*

Number	Square	Square root	Number	Square	Square root
1	1	1 0000	41	16 81	6 4031
2	4	1 4142	42	17 64	6 4807
3	9	1 7321	43	18 49	6 5574
4	16	2 0000	44	19 36	6 6332
5	25	2 2361	45	20 25	6 7082
6	36	2 4495	46	21 16	6 7823
7	49	2 6458	47	22 09	6 8557
8	64	2 8284	48	23 04	6 9282
9	81	3 0000	49	24 01	7 0000
10	1 00	3 1623	50	25 00	7 0711
11	1 21	3 3166	51	26 01	7 1414
12	1 44	3 4641	52	27 04	7 2111
13	1 69	3 6056	53	28 09	7 2801
14	1 96	3 7417	54	29 16	7 3485
15	2 25	3 8730	55	30 25	7 4162
16	2 56	4 0000	56	31 36	7 4833
17	2 89	4 1231	57	32 49	7 5498
18	3 24	4 2426	58	33 64	7 6158
19	3 61	4 3589	59	34 81	7 6811
20	4 00	4 4721	60	36 00	7 7460
21	4 41	4 5826	61	37 21	7 8102
22	4 84	4 6904	62	38 44	7 8740
23	5 29	4 7958	63	39 69	7 9373
24	5 76	4 8990	64	40 96	8 0000
25	6 25	5 0000	65	42 25	8 0623
26	6 76	5 0990	66	43 56	8 1240
27	7 29	5 1962	67	44 89	8 1854
28	7 84	5 2915	68	46 24	8 2462
29	8 41	5 3852	69	47 61	8 3066
30	9 00	5 4772	70	49 00	8 3666
31	9 61	5 5678	71	50 41	8 4261
32	10 24	5 6569	72	51 84	8 4853
33	10 89	5 7446	73	53 29	8 5440
34	11 56	5 8310	74	54 76	8 6023
35	12 25	5 9161	75	56 25	8 6603
36	12 96	6 0000	76	57 76	8 7178
37	13 69	6 0828	77	59 29	8 7750
38	14 44	6 1644	78	60 84	8 8318
39	15 21	6 2450	79	62 41	8 8882
40	16 00	6 3246	80	64 00	8 9443

* From Sorenson. Statistics for students of psychology and education.

TABLE A.—SQUARES AND SQUARE ROOTS OF NUMBERS FROM 1 TO 1,000.*—
(Continued)

Number	Square	Square root	Number	Square	Square root
81	65 61	9 0000	121	1 46 41	11 0000
82	67 24	9 0554	122	1 48 84	11 0454
83	68 89	9 1104	123	1 51 29	11 0905
84	70 56	9 1652	124	1 53 76	11 1355
85	72 25	9 2195	125	1 56 25	11 1803
86	73 96	9 2736	126	1 58 76	11 2250
87	75 69	9 3274	127	1 61 29	11 2694
88	77 44	9 3808	128	1 63 84	11 3137
89	79 21	9 4340	129	1 66 41	11 3578
90	81 00	9 4868	130	1 69 00	11 4018
91	82 81	9 5394	131	1 71 61	11.4455
92	84 64	9 5917	132	1 74 24	11 4891
93	86 49	9 6437	133	1 76 89	11 5326
94	88 36	9 6954	134	1 79 56	11 5758
95	90 25	9 7468	135	1 82 25	11 6190
96	92 16	9 7980	136	1 84 96	11 6619
97	94 09	9 8489	137	1 87 69	11 7047
98	96 04	9 8995	138	1 90 44	11 7473
99	98 01	9 9499	139	1 93 21	11 7898
100	1 00 00	10 0000	140	1 96 00	11 8322
101	1 02 01	10 0499	141	1 98 81	11 8743
102	1 04 04	10 0995	142	2 01 64	11 9164
103	1 06 09	10 1489	143	2 04 49	11.9583
104	1 08 16	10 1980	144	2 07 36	12 0000
105	1 10 25	10 2470	145	2 10 25	12 0416
106	1 12 36	10 2956	146	2 13 16	12 0830
107	1 14 49	10 3441	147	2 16 09	12 1244
108	1 16 64	10 3923	148	2 19 04	12 1655
109	1 18 81	10 4403	149	2 22 01	12 2066
110	1 21 00	10 4881	150	2 25 00	12 2474
111	1 23 21	10 5357	151	2 28 01	12 2882
112	1 25 44	10 5830	152	2 31 04	12 3288
113	1 27 69	10 6301	153	2 34 09	12 3693
114	1 29 96	10 6771	154	2 37 16	12 4097
115	1 32 25	10 7238	155	2 40 25	12.4499
116	1 34 56	10 7703	156	2 43 36	12 4900
117	1 36 89	10.8167	157	2 46 49	12.5300
118	1 39 24	10 8628	158	2 49 64	12 5698
119	1 41 61	10 9087	159	2 52 81	12 6095
120	1 44 00	10 9545	160	2 56 00	12.6491

* From Sorenson. Statistics for students of psychology and education.

TABLE A.—SQUARES AND SQUARE ROOTS OF NUMBERS FROM 1 TO 1,000 *—
(Continued)

Number	Square	Square root	Number	Square	Square root
161	2 59 21	12 6886	201	4 04 01	14 1774
162	2 62 44	12 7279	202	4 08 04	14 2127
163	2 65 69	12 7671	203	4 12 09	14 2478
164	2 68 96	12 8062	204	4 16 16	14 2829
165	2 72 25	12 8452	205	4 20 25	14 3178
166	2 75 56	12 8841	206	4 24 36	14 3527
167	2 78 89	12 9228	207	4 28 49	14 3875
168	2 82 24	12 9615	208	4 32 64	14 4222
169	2 85 61	13 0000	209	4 36 81	14 4568
170	2 89 00	13 0384	210	4 41 00	14 4914
171	2 92 41	13 0767	211	4 45 21	14 5258
172	2 95 84	13 1149	212	4 49 44	14 5602
173	2 99 29	13 1529	213	4 53 69	14 5945
174	3 02 76	13 1909	214	4 57 96	14 6287
175	3 06 25	13 2288	215	4 62 25	14 6629
176	3 09 76	13 2665	216	4 66 56	14 6969
177	3 13 29	13 3041	217	4 70 89	14 7309
178	3 16 84	13 3417	218	4 75 24	14 7648
179	3 20 41	13 3791	219	4 79 61	14 7986
180	3 24 00	13 4164	220	4 84 00	14 8324
181	3 27 61	13 4536	221	4 88 41	14 8661
182	3 31 24	13 4907	222	4 92 84	14 8997
183	3 34 89	13 5277	223	4 97 29	14 9332
184	3 38 56	13 5647	224	5 01 76	14 9666
185	3 42 25	13 6015	225	5 06 25	15 0000
186	3 45 96	13 6382	226	5 10 76	15 0333
187	3 49 69	13 6748	227	5 15 29	15 0665
188	3 53 44	13 7113	228	5 19 84	15 0997
189	3 57 21	13 7477	229	5 24 41	15 1327
190	3 61 00	13 7840	230	5 29 00	15 1658
191	3 64 81	13 8203	231	5 33 61	15 1987
192	3 68 64	13 8564	232	5 38 24	15 2315
193	3 72 49	13 8924	233	5 42 89	15 2643
194	3 76 36	13 9284	234	5 47 56	15 2971
195	3 80 25	13 9642	235	5 52 25	15 3297
196	3 84 16	14 0000	236	5 56 96	15 3623
197	3 88 09	14 0357	237	5 61 69	15 3948
198	3 92 04	14 0712	238	5 66 44	15 4272
199	3 96 01	14 1067	239	5 71 21	15 4596
200	4 00 00	14 1421	240	5 76 00	15 4919

* From Sorenson. Statistics for students of psychology and education.

TABLE A.—SQUARES AND SQUARE ROOTS OF NUMBERS FROM 1 TO 1,000.*—
(Continued)

Number	Square	Square root	Number	Square	Square root
241	5 80 81	15 5242	281	7 89 61	16 7631
242	5 85 64	15 5563	282	7 95 24	16 7929
243	5 90 49	15 5885	283	8 00 89	16 8226
244	5 95 36	15 6205	284	8 06 56	16 8523
245	6 00 25	15 6525	285	8 12 25	16.8819
246	6 05 16	15 6844	286	8 17 96	16 9115
247	6 10 09	15 7162	287	8 23 69	16 9411
248	6 15 04	15 7480	288	8 29 44	16 9706
249	6 20 01	15 7797	289	8 35 21	17 0000
250	6 25 00	15 8114	290	8 41 00	17 0294
251	6 30 01	15 8430	291	8 46 81	17 0587
252	6 35 04	15 8745	292	8 52 64	17 0880
253	6 40 09	15 9060	293	8 58 49	17 1172
254	6 45 16	15 9374	294	8 64 36	17 1464
255	6 50 25	15 9687	295	8 70 25	17 1756
256	6 55 36	16 0000	296	8 76 16	17 2047
257	6 60 49	16 0312	297	8 82 09	17.2337
258	6 65 64	16 0624	298	8 88 04	17 2627
259	6 70 81	16 0935	299	8 94 01	17.2916
260	6 76 00	16.1245	300	9 00 00	17 3205
261	6 81 21	16 1555	301	9 06 01	17 3494
262	6 86 44	16 1864	302	9 12 04	17 3781
263	6 91 69	16 2173	303	9 18 09	17 4069
264	6 96 96	16 2481	304	9 24 16	17 4356
265	7 02 25	16 2788	305	9 30 25	17 4642
266	7 07 56	16 3095	306	9 36 36	17 4929
267	7 12 89	16 3401	307	9 42 49	17 5214
268	7 18 24	16.3707	308	9 48 64	17 5499
269	7 23 61	16 4012	309	9 54 81	17 5784
270	7 29 00	16 4317	310	9 61 00	17.6068
271	7 34 41	16 4621	311	9 67 21	17 6352
272	7 39 84	16 4924	312	9 73 44	17 6635
273	7 45 29	16.5227	313	9 79 69	17 6918
274	7 50 76	16 5529	314	9 85 96	17.7200
275	7 56 25	16.5831	315	9 92 25	17 7482
276	7 61 76	16 6132	316	9 98 56	17 7764
277	7 67 29	16 6433	317	10 04 89	17.8045
278	7 72 84	16 6733	318	10 11 24	17.8326
279	7 78 41	16 7033	319	10 17 61	17 8606
280	7 84 00	16.7332	320	10 24 00	17.8885

* From Sorenson. Statistics for students of psychology and education.

TABLE A.—SQUARES AND SQUARE ROOTS OF NUMBERS FROM 1 TO 1,000.*—
(Continued)

Number	Square	Square root	Number	Square	Square root
321	10 30 41	17 9165	361	13 03 21	19 0000
322	10 36 84	17 9444	362	13 10 44	19 0263
323	10 43 29	17 9722	363	13 17 69	19 0526
324	10 49 76	18 0000	364	13 24 96	19 0788
325	10 56 25	18 0278	365	13 32 25	19 1050
326	10 62 76	18 0555	366	13 39 56	19 1311
327	10 69 29	18 0831	367	13 46 89	19 1572
328	10 75 84	18 1108	368	13 54 24	19 1833
329	10 82 41	18 1384	369	13 61 61	19 2094
330	10 89 00	18 1659	370	13 69 00	19 2354
331	10 95 61	18 1934	371	13 76 41	19 2614
332	11 02 24	18 2209	372	13 83 84	19 2873
333	11 08 89	18 2483	373	13 91 29	19 3132
334	11 15 56	18 2757	374	13 98 76	19 3391
335	11 22 25	18 3030	375	14 06 25	19 3649
336	11 28 96	18 3303	376	14 13 76	19 3907
337	11 35 69	18 3576	377	14 21 29	19 4165
338	11 42 44	18 3848	378	14 28 84	19 4422
339	11 49 21	18 4120	379	14 36 41	19 4679
340	11 56 00	18 4391	380	14 44 00	19 4936
341	11 62 81	18 4662	381	14 51 61	19 5192
342	11 69 64	18 4932	382	14 59 24	19 5448
343	11 76 49	18 5203	383	14 66 89	19 5704
344	11 83 36	18 5472	384	14 74 56	19 5959
345	11 90 25	18 5742	385	14 82 25	19 6214
346	11 97 16	18 6011	386	14 89 96	19 6469
347	12 04 09	18 6279	387	14 97 69	19 6723
348	12 11 04	18 6548	388	15 05 44	19 6977
349	12 18 01	18 6815	389	15 13 21	19 7231
350	12 25 00	18 7083	390	15 21 00	19 7484
351	12 32 01	18 7350	391	15 28 81	19 7737
352	12 39 04	18 7617	392	15 36 64	19 7990
353	12 46 09	18 7883	393	15 44 49	19 8242
354	12 53 16	18 8149	394	15 52 36	19 8494
355	12 60 25	18 8414	395	15 60 25	19 8746
356	12 67 36	18 8680	396	15 68 16	19 8997
357	12 74 49	18 8944	397	15 76 09	19 9249
358	12 81 64	18 9209	398	15 84 04	19 9499
359	12 88 81	18 9473	399	15 92 01	19 9750
360	12 96 00	18 9737	400	16 00 00	20 0000

* From Sorenson. Statistics for students of psychology and education.

TABLE A—SQUARES AND SQUARE ROOTS OF NUMBERS FROM 1 TO 1,000 *—
(Continued)

Number	Square	Square root	Number	Square	Square root
401	16 08 01	20 0250	441	19 44 81	21 0000
402	16 16 04	20 0499	442	19 53 64	21 0238
403	16 24 09	20 0749	443	19 62 49	21 0476
404	16 32 16	20 0998	444	19 71 36	21 0713
405	16 40 25	20 1246	445	19 80 25	21 0950
406	16 48 36	20 1494	446	19 89 16	21 1187
407	16 56 49	20 1742	447	19 98 09	21 1424
408	16 64 64	20 1990	448	20 07 04	21 1660
409	16 72 81	20 2237	449	20 16 01	21 1896
410	16 81 00	20 2485	450	20 25 00	21 2132
411	16 89 21	20 2731	451	20 34 01	21 2368
412	16 97 44	20 2978	452	20 43 04	21 2603
413	17 05 69	20 3224	453	20 52 09	21 2838
414	17 13 96	20 3470	454	20 61 16	21 3073
415	17 22 25	20 3715	455	20 70 25	21 3307
416	17 30 56	20 3961	456	20 79 36	21 3542
417	17 38 89	20 4206	457	20 88 49	21 3776
418	17 47 24	20 4450	458	20 97 64	21 4009
419	17 55 61	20 4695	459	21 06 81	21 4243
420	17 64 00	20 4939	460	21 16 00	21 4476
421	17 72 41	20 5183	461	21 25 21	21 4709
422	17 80 84	20 5426	462	21 34 44	21 4942
423	17 89 29	20 5670	463	21 43 69	21 5174
424	17 97 76	20 5913	464	21 52 96	21 5407
425	18 06 25	20 6155	465	21 62 25	21 5639
426	18 14 76	20 6398	466	21 71 56	21 5870
427	18 23 29	20 6640	467	21 80 89	21 6102
428	18 31 84	20 6882	468	21 90 24	21 6333
429	18 40 41	20 7123	469	21 99 61	21 6564
430	18 49 00	20 7364	470	22 09 00	21 6795
431	18 57 61	20 7605	471	22 18 41	21 7025
432	18 66 24	20 7846	472	22 27 84	21 7256
433	18 74 89	20 8087	473	22 37 29	21 7486
434	18 83 56	20 8327	474	22 46 76	21 7715
435	18 92 25	20 8567	475	22 56 25	21 7945
436	19 00 96	20 8806	476	22 65 76	21 8174
437	19 09 69	20 9045	477	22 75 29	21 8403
438	19 18 44	20 9284	478	22 84 84	21 8632
439	19 27 21	20 9523	479	22 94 41	21 8861
440	19 36 00	20 9762	480	23 04 00	21 9089

* From Sorenson. Statistics for students of psychology and education.

TABLE A.—SQUARES AND SQUARE ROOTS OF NUMBERS FROM 1 TO 1,000.*—
(Continued)

Number	Square	Square root	Number	Square	Square root
481	23 13 61	21 9317	521	27 14 41	22 8254
482	23 23 24	21 9545	522	27 24 84	22 8473
483	23 32 89	21 9773	523	27 35 29	22 8692
484	23 42 56	22 0000	524	27 45 76	22 8910
485	23 52 25	22 0227	525	27 56 25	22 9129
486	23 61 96	22 0454	526	27 66 76	22 9347
487	23 71 69	22 0681	527	27 77 29	22 9565
488	23 81 44	22 0907	528	27 87 84	22 9783
489	23 91 21	22 1133	529	27 98 41	23 0000
490	24 01 00	22 1359	530	28 09 00	23 0217
491	24 10 81	22 1585	531	28 19 61	23 0434
492	24 20 64	22 1811	532	28 30 24	23 0651
493	24 30 49	22 2036	533	28 40 89	23 0868
494	24 40 36	22 2261	534	28 51 56	23 1084
495	24 50 25	22 2486	535	28 62 25	23 1301
496	24 60 16	22 2711	536	28 72 96	23 1517
497	24 70 09	22 2935	537	28 83 69	23 1733
498	24 80 04	22 3159	538	28 94 44	23 1948
499	24 90 01	22 3383	539	29 05 21	23 2164
500	25 00 00	22 3607	540	29 16 00	23 2379
501	25 10 01	22 3830	541	29 26 81	23 2594
502	25 20 04	22 4054	542	29 37 64	23 2809
503	25 30 09	22 4277	543	29 48 49	23 3024
504	25 40 16	22 4499	544	29 59 36	23 3238
505	25 50 25	22 4722	545	29 70 25	23 3452
506	25 60 36	22 4944	546	29 81 16	23 3666
507	25 70 49	22 5167	547	29 92 09	23 3880
508	25 80 64	22 5389	548	30 03 04	23 4094
509	25 90 81	22 5610	549	30 14 01	23 4307
510	26 01 00	22 5832	550	30 25 00	23 4521
511	26 11 21	22 6053	551	30 36 01	23 4734
512	26 21 44	22 6274	552	30 47 04	23 4947
513	26 31 69	22 6495	553	30 58 09	23 5160
514	26 41 96	22 6716	554	30 69 16	23 5372
515	26 52 25	22 6936	555	30 80 25	23 5584
516	26 62 56	22 7156	556	30 91 36	23 5797
517	26 72 89	22 7376	557	31 02 49	23 6008
518	26 83 24	22 7596	558	31 13 64	23 6220
519	26 93 61	22 7816	559	31 24 81	23 6432
520	27 04 00	22 8035	560	31 36 00	23 6643

* From Sorenson. Statistics for students of psychology and education.

TABLE A.—SQUARES AND SQUARE ROOTS OF NUMBERS FROM 1 TO 1,000 *—
(Continued)

Number	Square	Square root	Number	Square	Square root
561	31 47 21	23 6854	601	36 12 01	24 5153
562	31 58 44	23 7065	602	36 24 04	24 5357
563	31 69 69	23 7276	603	36 36 09	24 5561
564	31 80 96	23 7487	604	36 48 16	24 5764
565	31 92 25	23 7697	605	36 60 25	24 5967
566	32 03 56	23 7908	606	36 72 36	24 6171
567	32 14 89	23 8118	607	36 84 49	24 6374
568	32 26 24	23 8328	608	36 96 64	24 6577
569	32 37 61	23 8537	609	37 08 81	24 6779
570	32 49 00	23 8747	610	37 21 00	24 6982
571	32 60 41	23 8956	611	37 33 21	24.7184
572	32 71 84	23 9165	612	37 45 44	24.7385
573	32 83 29	23 9374	613	37 57 69	24 7588
574	32 94 76	23 9583	614	37 69 96	24 7790
575	33 06 25	23 9792	615	37 82 25	24 7992
576	33 17 76	24 0000	616	37 94 56	24 8193
577	33 29 29	24 0208	617	38 06 89	24 8395
578	33 40 84	24 0416	618	38 19 24	24 8596
579	33 52 41	24 0624	619	38 31 61	24 8797
580	33 64 00	24 0832	620	38 44 00	24 8998
581	33 75 61	24 1039	621	38 56 41	24.9199
582	33 87 24	24 1247	622	38 68 84	24 9399
583	33 98 89	24 1454	623	38 81 29	24 9600
584	34 10 56	24 1661	624	38 93 76	24 9800
585	34 22 25	24 1868	625	39 06 25	25 0000
586	34 33 96	24 2074	626	39 18 76	25 0200
587	34 45 69	24 2281	627	39 31 29	25.0400
588	34 57 44	24 2487	628	39 43 84	25 0599
589	34 69 21	24 2693	629	39 56 41	25 0799
590	34 81 00	24.2899	630	39 69 00	25.0998
591	34 92 81	24 3105	631	39 81 61	25.1197
592	35 04 64	24 3311	632	39 94 24	25.1396
593	35 16 49	24 3516	633	40 06 89	25 1595
594	35 28 36	24 3721	634	40 19 56	25 1794
595	35 40 25	24 3926	635	40 32 25	25.1992
596	35 52 16	24.4131	636	40 44 96	25 2190
597	35 64 09	24 4336	637	40 57 69	25 2389
598	35 76 04	24 4540	638	40 70 44	25 2587
599	35 88 01	24 4745	639	40 83 21	25.2784
600	36 00 00	24 4949	640	40 96 00	25 2982

* From Sorenson Statistics for students of psychology and education.

TABLE A—SQUARES AND SQUARE ROOTS OF NUMBERS FROM 1 TO 1,000.*—
(Continued)

Number	Square	Square root	Number	Square	Square root
641	41 08 81	25 3180	681	46 37 61	26 0960
642	41 21 64	25 3377	682	46 51 24	26 1151
643	41 34 49	25 3574	683	46 64 89	26 1343
644	41 47 36	25 3772	684	46 78 56	26 1534
645	41 60 25	25 3969	685	46 92 25	26 1725
646	41 73 16	25 4165	686	47 05 96	26 1916
647	41 86 09	25 4362	687	47 19 69	26 2107
648	41 99 04	25 4558	688	47 33 44	26 2298
649	42 12 01	25 4755	689	47 47 21	26 2488
650	42 25 00	25 4951	690	47 61 00	26 2679
651	42 38 01	25 5147	691	47 74 81	26 2869
652	42 51 04	25 5343	692	47 88 64	26 3059
653	42 64 09	25 5539	693	48 02 49	26 3249
654	42 77 16	25 5734	694	48 16 36	26 3439
655	42 90 25	25 5930	695	48 30 25	26 3629
656	43 03 36	25 6125	696	48 44 16	26 3818
657	43 16 49	25 6320	697	48 58 09	26 4008
658	43 29 64	25 6515	698	48 72 04	26 4197
659	43 42 81	25 6710	699	48 86 01	26 4386
660	43 56 00	25 6905	700	49 00 00	26 4575
661	43 69 21	25 7099	701	49 14 01	26 4764
662	43 82 44	25 7294	702	49 28 04	26 4953
663	43 95 69	25 7488	703	49 42 09	26 5141
664	44 08 96	25 7682	704	49 56 16	26 5330
665	44 22 25	25 7876	705	49 70 25	26 5518
666	44 35 56	25 8070	706	49 84 36	26 5707
667	44 48 89	25 8263	707	49 98 49	26 5895
668	44 62 24	25 8457	708	50 12 64	26 6083
669	44 75 61	25 8650	709	50 26 81	26 6271
670	44 89 00	25 8844	710	50 41 00	26 6458
671	45 02 41	25 9037	711	50 55 21	26 6646
672	45 15 84	25 9230	712	50 69 44	26 6833
673	45 29 29	25 9422	713	50 83 69	26 7021
674	45 42 76	25 9615	714	50 97 96	26 7208
675	45 56 25	25 9808	715	51 12 25	26 7395
676	45 69 76	26 0000	716	51 26 56	26 7582
677	45 83 29	26 0192	717	51 40 89	26 7769
678	45 96 84	26 0384	718	51 55 24	26 7955
679	46 10 41	26 0576	719	51 69 61	26 8142
680	46 24 00	26 0768	720	51 84 00	26 8328

* From Sorenson. Statistics for students of psychology and education.

TABLE A—SQUARES AND SQUARE ROOTS OF NUMBERS FROM 1 TO 1,000 *—
(Continued)

Number	Square	Square root	Number	Square	Square root
721	51 98 41	26 8514	761	57 91 21	27 5862
722	52 12 84	26 8701	762	58 06 44	27 6043
723	52 27 29	26 8887	763	58 21 69	27 6225
724	52 41 76	26 9072	764	58 36 96	27 6405
725	52 56 25	26 9258	765	58 52 25	27 6586
726	52 70 76	26 9444	766	58 67 56	27 6767
727	52 85 29	26 9629	767	58 82 89	27 6948
728	52 99 84	26 9815	768	58 98 24	27 7128
729	53 14 41	27 0000	769	59 13 61	27 7308
730	53 29 00	27 0185	770	59 29 00	27 7489
731	53 43 61	27 0370	771	59 44 41	27 7669
732	53 58 24	27 0555	772	59 59 84	27 7849
733	53 72 89	27 0740	773	59 75 29	27 8029
734	53 87 56	27 0924	774	59 90 76	27 8209
735	54 02 25	27 1109	775	60 06 25	27 8388
736	54 16 96	27 1293	776	60 21 76	27 8568
737	54 31 69	27 1477	777	60 37 29	27 8747
738	54 46 44	27 1662	778	60 52 84	27 8927
739	54 61 27	27 1846	779	60 68 41	27 9106
740	54 76 00	27 2029	780	60 84 00	27 9285
741	54 90 81	27 2213	781	60 99 61	27 9464
742	55 05 64	27 2397	782	61 15 24	27 9643
743	55 20 49	27 2580	783	61 30 89	27 9821
744	55 35 36	27 2764	784	61 46 56	28 0000
745	55 50 25	27 2947	785	61 62 25	28 0179
746	55 65 16	27 3130	786	61 77 96	28 0357
747	55 80 09	27 3313	787	61 93 69	28 0535
748	55 95 04	27 3496	788	62 09 44	28 0713
749	56 10 01	27 3679	789	62 25 21	28 0891
750	56 25 00	27 3861	790	62 41 00	28 1069
751	56 40 01	27 4044	791	62 56 81	28 1247
752	56 55 04	27 4226	792	62 72 64	28 1425
753	56 70 09	27 4408	793	62 88 49	28 1603
754	56 85 16	27 4591	794	63 04 36	28 1780
755	57 00 25	27 4773	795	63 20 25	28 1957
756	57 15 36	27 4955	796	63 36 16	28 2135
757	57 30 49	27 5136	797	63 52 09	28 2312
758	57 45 64	27 5318	798	63 68 04	28 2489
759	57 60 81	27 5500	799	63 84 01	28 2666
760	57 76 00	27 5681	800	64 00 00	28 2843

* From Sorenson. Statistics for students of psychology and education.

TABLE A—SQUARES AND SQUARE ROOTS OF NUMBERS FROM 1 TO 1,000 *—
(Continued)

Number	Square	Square root	Number	Square	Square root
801	64 16 01	28.3019	841	70 72 81	29 0000
802	64 32 04	28 3196	842	70 89 64	29 0172
803	64 48 09	28 3373	843	71 06 49	29 0345
804	64 64 16	28 3049	844	71 23 36	29 0517
805	64 80 25	28 3725	845	71 40 25	29 0689
806	64 96 36	28 3901	846	71 57 16	29 0861
807	65 12 49	28 4077	847	71 74 09	29 1033
808	65 28 64	28 4253	848	71 91 04	29 1204
809	65 44 81	28 4429	849	72 08 01	29 1376
810	65 61 00	28 4605	850	72 25 00	29.1548
811	65 77 21	28 4781	851	72 42 01	29 1719
812	65 93 44	28 4956	852	72 59 04	29 1890
813	66 09 69	28 5132	853	72 76 09	29 2062
814	66 25 96	28 5307	854	72 93 16	29 2233
815	66 42 25	28.5482	855	73 10 25	29 2404
816	66 58 56	28 5657	856	73 27 36	29 2575
817	66 74 89	28 5832	857	73 44 49	29 2746
818	66 91 24	28 6007	858	73 61 64	29 2916
819	67 07 61	28 6082	859	73 78 81	29 3087
820	67 24 00	28 6356	860	73 96 00	29 3258
821	67 40 41	28 6531	861	74 13 21	29 3428
822	67 56 84	28 6705	862	74 30 44	29 3598
823	67 73 29	28 6880	863	74 47 69	29 3769
824	67 89 76	28 7054	864	74 64 96	29 3939
825	68 06 25	28 7228	865	74 82 25	29 4109
826	68 22 76	28 7402	866	74 99 56	29 4279
827	68 39 29	28 7576	867	75 16 89	29 4449
828	68 55 84	28 7750	868	75 34 24	29 4618
829	68 72 41	28.7924	869	75 51 61	29 4788
830	68 89 00	28 8097	870	75 69 00	29 4958
831	69 05 61	28 8271	871	75 86 41	29 5127
832	69 22 24	28 8444	872	76 03 84	29 5296
833	69 38 89	28 8617	873	76 21 29	29 5466
834	69 55 56	28 8791	874	76 38 76	29 5635
835	69 72 25	28 8964	875	76 56 25	29 5804
836	69 88 96	28 9137	876	76 73 76	29 5973
837	70 05 69	28 9310	877	76 91 29	29 6142
838	70 22 44	28 9482	878	77 08 84	29 6311
839	70 39 21	28 9655	879	77 26 41	29 6479
840	70 56 00	28 9828	880	77 44 00	29 6648

* From Sorenson. Statistics for students of psychology and education.

TABLE A—SQUARES AND SQUARE ROOTS OF NUMBERS FROM 1 TO 1,000 *—
(Continued)

Number	Square	Square root	Number	Square	Square root
881	77 61 61	29 6816	921	84 82 41	30 3480
882	77 79 24	29 6985	922	85 00 84	30 3645
883	77 96 89	29 7153	923	85 19 29	30 3809
884	78 14 56	29 7321	924	85 37 76	30 3974
885	78 32 25	29 7489	925	85 56 25	30 4138
886	78 49 96	29 7658	926	85 74 76	30 4302
887	78 67 69	29 7825	927	85 93 29	30 4467
888	78 85 44	29 7993	928	86 11 84	30 4631
889	79 03 21	29 8161	929	86 30 41	30 4795
890	79 21 00	29 8329	930	86 49 00	30 4959
891	79 38 81	29 8496	931	86 67 61	30 5123
892	79 56 64	29 8664	932	86 86 24	30 5287
893	79 74 49	29 8831	933	87 04 89	30 5450
894	79 92 36	29 8998	934	87 23 56	30 5614
895	80 10 25	29.9166	935	87 42 25	30 5778
896	80 28 16	29 9333	936	87 60 96	30 5941
897	80 46 09	29 9500	937	87 79 69	30 6105
898	80 64 04	29 9666	938	87 98 44	30 6268
899	80 82 01	29 9833	939	88 17 21	30 6431
900	81 00 00	30.0000	940	88 36 00	30 6594
901	81 18 01	30 0167	941	88 54 81	30 6757
902	81 36 04	30 0333	942	88 73 64	30 6920
903	81 54 09	30 0500	943	88 92 49	30 7083
904	81 72 16	30.0666	944	89 11 36	30 7246
905	81 90 25	30 0832	945	89 30 25	30 7409
906	82 08 36	30.0998	946	89 49 16	30 7571
907	82 26 49	30 1164	947	89 68 09	30 7734
908	82 44 64	30 1330	948	89 87 04	30 7896
909	82 62 81	30 1496	949	90 06 01	30 8058
910	82 81 00	30 1662	950	90 25 00	30 8221
911	82 99 21	30.1828	951	90 44 01	30 8383
912	83 17 44	30 1993	952	90 63 04	30 8545
913	83 35 69	30 2159	953	90 82 09	30 8707
914	83 53 96	30 2324	954	91 01 16	30 8869
915	83 72 25	30 2490	955	91 20 25	30 9031
916	83 90 56	30 2655	956	91 39 36	30.9192
917	84 08 89	30.2820	957	91 58 49	30.9354
918	84 27 24	30 2985	958	91 77 64	30 9516
919	84 45 61	30 3150	959	91 96 81	30 9677
920	84 64 00	30.3315	960	92 16 00	30 9839

* From Sorenson. Statistics for students of psychology and education

TABLE A.—SQUARES AND SQUARE ROOTS OF NUMBERS FROM 1 TO 1,000.*—
(Continued)

Number	Square	Square root	Number	Square	Square root
961	92 35 21	31.0000	981	96 23 61	31 3209
962	92 54 44	31.0161	982	96 43 24	31 3369
963	92 73 69	31 0322	983	96 62 89	31 3528
964	92 92 96	31 0483	984	96 82 56	31 3688
965	93 12 25	31 0644	985	97 02 25	31 3847
966	93 31 56	31 0805	986	97 21 96	31 4006
967	93 50 89	31 0966	987	97 41 69	31 4166
968	93 70 24	31 1127	988	97 61 44	31 4325
969	93 89 61	31 1288	989	97 81 21	31 4484
970	94 09 00	31.1448	990	98 01 00	31 4643
971	94 28 41	31.1609	991	98 20 81	31 4802
972	94 47 84	31 1769	992	98 40 64	31 4960
973	94 67 29	31 1929	993	98 60 49	31.5119
974	94 86 76	31 2090	994	98 80 36	31 5278
975	95 06 25	31 2250	995	99 00 25	31 5436
976	95 25 76	31.2410	996	99 20 16	31 5595
977	95 45 29	31 2570	997	99 40 09	31.5753
978	95 64 84	31.2730	998	99 60 04	31 5911
979	95 84 41	31 2890	999	99 80 01	31.6070
980	96 04 00	31.3050	1000	100 00 00	31 6228

* From Sorenson, H. Statistics for students of psychology and education. New York. McGraw-Hill, 1936.

TABLE B.—PROPORTIONS OF THE AREA UNDER THE NORMAL DISTRIBUTION CURVE AND ORDINATES CORRESPONDING TO GIVEN STANDARD SCORES

<i>z</i>	<i>A</i>	<i>B</i>	<i>C</i>	<i>y</i>
Standard score (x/σ)	Area from mean to x/σ	Area in larger portion	Area in smaller portion	Ordinate at x/σ
0 00	0000	.5000	5000	3989
0 05	.0199	.5199	.4801	3984
0 10	.0398	.5398	.4602	.3970
0.15	.0596	.5596	.4404	.3945
0.20	.0793	.5793	.4207	.3910
0 25	.0987	.5987	.4013	3867
0 30	.1179	.6179	.3821	3814
0.35	.1368	.6368	.3632	3752
0 40	.1554	.6554	.3446	3683
0.45	.1736	.6736	.3264	3605
0 50	.1915	.6915	.3085	.3521
0.55	.2088	.7088	.2912	.3429
0 60	.2257	.7257	.2743	.3332
0.65	.2422	.7422	.2578	.3230
0.70	.2580	.7580	.2420	.3123
0.75	.2734	.7734	.2266	3011
0.80	.2881	.7881	.2119	.2897
0.85	.3023	.8023	.1977	2780
0.90	.3159	.8159	.1841	2661
0.95	.3289	.8289	.1711	.2541
1.00	.3413	.8413	.1587	.2420
1.05	.3531	.8531	.1469	2299
1.10	.3643	.8643	.1357	.2179
1 15	.3749	.8749	.1251	.2059
1.20	.3849	.8849	.1151	1942
1.25	.3944	.8944	.1056	1826
1 30	.4032	.9032	.0968	1714
1.35	.4115	.9115	.0885	1604
1 40	.4192	.9192	.0808	1497
1.45	.4265	.9265	.0735	.1394
1 50	.4332	.9332	.0668	.1295
1.55	.4394	.9394	.0606	.1200
1 60	.4452	.9452	.0548	.1109
1.65	.4505	.9505	.0495	.1023
1.70	.4554	.9554	.0446	.0940

TABLE B.—PROPORTIONS OF THE AREA UNDER THE NORMAL DISTRIBUTION CURVE AND ORDINATES CORRESPONDING TO GIVEN STANDARD SCORES —(Continued)

z	A	B	C	y
Standard score (x/σ)	Area from mean to x/σ	Area in larger portion	Area in smaller portion	Ordinate at x/σ
1 75	4599	9599	0401	0863
1 80	4641	9641	0359	0790
1 85	4678	9678	0322	0721
1.90	4713	9713	0287	0656
1 95	.4744	9744	0256	0596
2 00	4772	.9772	0228	0540
2 05	4798	.9798	0202	0488
2 10	4821	9821	0179	0440
2.15	4842	9842	0158	0395
2 20	4861	.9861	.0139	.0355
2 25	4878	9878	0122	0317
2 30	4893	.9893	.0107	0283
2 35	4906	9906	0094	0252
2.40	4918	.9918	.0082	0224
2.45	4929	.9929	.0071	.0198
2 50	.4938	9938	0062	.0175
2.55	4946	9946	0054	0154
2 60	4953	9953	0047	.0136
2 65	4960	9960	.0040	0119
2.70	4965	9965	0035	.0104
2 80	.4974	9974	0026	0079
2 90	4981	9981	0019	0060
3 00	.49865	99865	00135	0044
3.10	49903	99903	00097	0033
3 20	49931	.99931	00069	0024
3 40	.49966	99966	00034	0012
3 60	.49984	99984	00016	0006
3.80	49993	99993	00007	0003
4 00	4999683	.9999683	0000317	.0001
4.50	.4999966	.9999966	.0000034	000015
5.00	4999997	.9999997	0000003	0000016
6.00	.49999999	99999999	.000000001	000000006

TABLE C—STANDARD SCORES (OR DEVIATES) AND ORDINATES CORRESPONDING TO POINTS OF DIVISION OF THE AREA UNDER THE NORMAL CURVE INTO A LARGER PROPORTION (B) AND A SMALLER PROPORTION (C), ALSO THE VALUE \sqrt{BC}

B The larger area	z Standard score	y Ordinate	\sqrt{BC}	C The smaller area
.500	0000	3989	.5000	.500
.505	0125	3989	.5000	.495
.510	0251	3988	.4999	.490
.515	.0376	3987	.4998	.485
.520	.0502	.3984	.4996	.480
.525	.0627	3982	.4994	.475
.530	0753	3978	.4991	.470
.535	0878	3974	.4988	.465
.540	.1004	3969	.4984	.460
.545	1130	3964	.4980	.455
.550	1257	.3958	.4975	.450
.555	1383	3951	.4970	.445
.560	1510	3944	.4964	.440
.565	1637	3936	.4958	.435
.570	1764	3928	.4951	.430
.575	1891	3919	.4943	.425
.580	2019	3909	.4936	.420
.585	2147	3899	.4927	.415
.590	2275	3887	.4918	.410
.595	2404	3876	.4909	.405
.600	.2533	3863	.4899	.400
.605	2663	3850	.4889	.395
.610	2793	3837	.4877	.390
.615	2924	3822	.4867	.385
.620	3055	3808	.4854	.380
.625	3186	3792	.4841	.375
.630	3319	3776	.4828	.370
.635	3451	3759	.4814	.365
.640	.3585	3741	.4800	.360
.645	.3719	.3723	.4785	.355
.650	3853	.3704	.4770	.350
.655	3989	.3684	.4754	.345
.660	.4125	.3664	.4737	.340
.665	.4261	.3643	.4720	.335
.670	.4399	.3621	.4702	.330
.675	.4538	.3599	.4684	.325
.680	.4677	.3576	.4665	.320
.685	.4817	.3552	.4645	.315
.690	.4959	.3528	.4625	.310
.695	.5101	.3503	.4604	.305
.700	.5244	.3477	.4583	.300
.705	.5388	.3450	.4560	.295
.710	.5534	.3423	.4538	.290
.715	.5681	3395	.4514	.285
.720	5828	3366	.4490	.280

TABLE C—STANDARD SCORES (OR DEVIATES) AND ORDINATES CORRESPONDING TO POINTS OF DIVISION OF THE AREA UNDER THE NORMAL CURVE INTO A LARGER PROPORTION (B) AND A SMALLER PROPORTION (C), ALSO THE VALUE \sqrt{BC} —(Continued)

B The larger area	z Standard score	y Ordinate	\sqrt{BC}	C The smaller area
.725	5978	3337	4465	.275
.730	6128	3306	4440	.270
.735	6280	3275	4413	.265
.740	6433	3244	4386	.260
.745	6588	3211	4359	.255
.750	6745	3178	4330	.250
.755	6903	3144	4301	.245
.760	7063	3109	4271	.240
.765	7225	3073	4240	.235
.770	7388	3036	4208	.230
.775	7554	2999	4176	.225
.780	7722	2961	4142	.220
.785	7892	2922	4108	.215
.790	8064	2882	4073	.210
.795	8239	2841	4037	.205
.800	8416	2800	4000	.200
.805	8596	2757	3962	.195
.810	8779	2714	3923	.190
.815	8965	2669	3883	.185
.820	9154	2624	3842	.180
.825	.9346	2578	3800	.175
.830	.9542	2531	.3756	.170
.835	.9741	2482	3712	.165
.840	.9945	2433	3666	.160
.845	1 0152	.2383	3619	.155
.850	1 0364	2332	3571	.150
.855	1 0581	2279	3521	.145
.860	1 0803	2226	3470	.140
.865	1 1031	2171	3417	.135
.870	1 1264	.2115	3363	.130
.875	1.1503	.2059	.3307	.125
.880	1 1750	2000	3250	.120
.885	1 2004	1941	3190	.115
.890	1 2265	1880	3129	.110
.895	1 2536	.1818	3066	.105
.900	1 2816	1755	3000	.100
.905	1 3016	1690	2932	.095
.910	1 3408	1624	2862	.090
.915	1 3722	1556	2789	.085
.920	1 4051	.1487	2713	.080
.925	1 4395	1416	2634	.075
.930	1 4757	1343	2551	.070
.935	1 5141	1268	2465	.065
.940	1.5548	1191	2375	.060
.945	1 5982	1112	2280	.055

TABLE C—STANDARD SCORES (OR DEVIATES) AND ORDINATES CORRESPONDING TO POINTS OF DIVISION OF THE AREA UNDER THE NORMAL CURVE INTO A LARGER PROPORTION (*B*) AND A SMALLER PROPORTION (*C*), ALSO THE VALUE \sqrt{BC} —(Continued)

<i>B</i> The larger area	<i>z</i> Standard score	<i>y</i> Ordinate	\sqrt{BC}	<i>C</i> The smaller area
.950	1 6449	1031	2179	050
.955	1 6954	0948	2073	045
.960	1 7507	0862	1960	040
.965	1 8119	0773	1838	035
.970	1 8808	0680	1706	030
.975	1 9600	0584	1561	025
.980	2 0537	0484	1400	020
.985	2 1701	0379	1216	015
.990	2 3263	0267	1411	.010
.995	2 5758	0145	0705	005
.996	2 6521	0118	0631	.004
.997	2 7478	0091	0547	003
.998	2 8782	0063	0447	002
.999	3 0902	0034	0316	001
.9995	3 2905	0018	0224	0005

TABLE D.—COEFFICIENTS OF CORRELATION AND *t* RATIOS SIGNIFICANT AT THE 5 PER CENT LEVEL (ROMAN TYPE) AND AT THE 1 PER CENT LEVEL (BOLD-FACED TYPE) FOR VARYING DEGREES OF FREEDOM*

Degrees of freedom	Number of variables									<i>t</i>
	2	3	4	5	6	7	9	13	25	
1	.997 1.000	.999 1.000	.999 1.000	.999 1.000	1.000 1.000	1.000 1.000	1.000 1.000	1.000 1.000	1.000 1.000	12 706 63.657
2	.950 .990	.975 .995	.983 .997	.987 .998	.990 .998	.992 .998	.994 .999	.996 .999	.998 1.000	4 303 9.925
3	.878 .969	.930 .976	.950 .983	.961 .987	.968 .990	.973 .991	.979 .993	.986 .995	.993 .998	3 182 5.841
4	.811 .917	.881 .949	.912 .962	.930 .970	.942 .975	.950 .979	.961 .984	.973 .989	.986 .994	2 776 4.604
5	.754 .874	.836 .917	.874 .937	.898 .949	.914 .957	.925 .963	.941 .971	.958 .980	.978 .989	2 571 4.032
6	.707 .834	.795 .886	.839 .911	.867 .927	.886 .938	.900 .946	.920 .957	.943 .969	.969 .983	2 447 3.707
7	.666 .798	.758 .855	.807 .885	.838 .904	.860 .918	.876 .928	.900 .942	.927 .958	.960 .977	2 365 3.499
8	.632 .765	.726 .827	.777 .860	.811 .882	.835 .898	.854 .909	.880 .926	.912 .946	.950 .970	2 306 3.355
9	.602 .735	.697 .800	.750 .836	.786 .861	.812 .878	.832 .891	.861 .911	.897 .934	.941 .963	2 262 3.250
10	.576 .708	.671 .776	.726 .814	.763 .840	.790 .859	.812 .874	.843 .895	.882 .922	.932 .955	2 228 3.169
11	.553 .684	.648 .753	.703 .793	.741 .821	.770 .841	.792 .857	.826 .880	.868 .910	.922 .948	2 201 3.106
12	.532 .661	.627 .732	.683 .773	.722 .802	.751 .824	.774 .841	.809 .866	.854 .898	.913 .940	2 179 3.055
13	.514 .641	.608 .712	.664 .755	.703 .785	.733 .807	.757 .825	.794 .852	.840 .886	.904 .932	2 160 3.012
14	.497 .623	.590 .694	.646 .737	.686 .768	.717 .792	.741 .810	.779 .838	.828 .875	.895 .924	2 145 2.977
15	.482 .606	.574 .677	.630 .721	.670 .752	.701 .776	.726 .796	.765 .825	.815 .864	.886 .917	2 131 2.947
16	.468 .590	.559 .662	.615 .706	.655 .738	.686 .762	.712 .782	.751 .813	.803 .853	.878 .909	2 120 2.921
17	.456 .576	.545 .647	.601 .691	.641 .724	.673 .749	.698 .769	.738 .800	.792 .842	.869 .902	2 110 2.898
18	.444 .561	.532 .633	.587 .678	.628 .710	.660 .736	.686 .756	.726 .789	.781 .832	.861 .894	2 101 2.878
19	.433 .549	.520 .620	.575 .665	.615 .698	.647 .723	.674 .744	.714 .778	.770 .822	.853 .887	2 093 2.861
20	.423 .537	.509 .608	.563 .652	.604 .685	.636 .712	.662 .733	.703 .767	.760 .812	.845 .880	2 086 2.845
21	.413 .526	.498 .596	.552 .641	.592 .674	.624 .700	.651 .722	.693 .756	.750 .803	.837 .873	2 080 2.831
22	.404 .515	.488 .585	.542 .630	.582 .663	.614 .690	.640 .712	.682 .746	.740 .794	.830 .866	2 074 2.819
23	.396 .508	.479 .574	.532 .619	.572 .652	.604 .679	.630 .701	.673 .736	.731 .785	.823 .859	2 069 2.807

* Adapted from Wallace, H. A., and Snedecor, G. W., Correlation and machine calculation, 1931, by courtesy of the authors.

TABLE E.—TABLE OF CHI SQUARE*

#	P = .99	.98	.95	.90	.80	.70	.50	.30	.20	.10	.05	.02	.01
1	.000157	.000628	.00393	.0158	.0642	.148	.455	1.074	1.642	2.706	3.841	5.412	6.635
2	.0201	.0404	.103	.211	.446	.713	1.386	2.408	3.219	4.605	5.991	7.824	9.210
3	.115	.185	.352	.584	1.005	1.424	2.366	3.665	4.642	6.251	7.815	9.837	11.341
4	.297	.439	.711	1.064	1.649	2.195	3.337	4.878	5.989	7.779	9.488	11.668	13.277
5	.554	.752	1.145	1.610	2.343	3.000	4.351	6.064	7.289	9.236	11.070	13.388	15.086
6	.872	1.134	1.635	2.204	3.070	3.828	5.348	7.231	8.558	10.645	12.592	15.033	16.812
7	1.239	1.564	2.167	2.833	3.822	4.671	6.346	8.383	9.803	12.017	14.067	16.222	18.475
8	1.646	2.032	2.733	3.490	4.594	5.537	7.344	9.524	11.030	13.362	15.507	18.168	20.090
9	2.088	2.532	3.325	4.168	5.380	6.339	8.342	10.656	12.242	14.684	16.919	19.679	21.666
10	2.558	3.059	3.940	4.865	6.179	7.267	9.342	11.781	13.442	15.987	18.307	21.161	23.209
11	3.053	3.609	4.575	5.578	6.989	8.148	10.341	12.899	14.631	17.275	19.675	22.618	24.725
12	3.571	4.178	5.226	6.304	7.807	9.034	11.340	14.011	15.812	18.549	21.026	24.054	26.217
13	4.107	4.765	5.892	7.042	8.634	9.926	12.340	15.119	16.985	19.812	22.362	25.472	27.688
14	4.660	5.368	6.571	7.790	9.467	10.821	13.339	16.222	18.151	21.064	23.685	26.873	29.141
15	5.229	5.985	7.261	8.547	10.307	11.721	14.339	17.322	19.311	22.307	24.996	28.259	30.578
16	5.812	6.614	7.962	9.312	11.152	12.624	15.338	18.418	20.465	23.542	26.296	29.633	32.000
17	6.408	7.255	8.672	10.085	12.002	13.531	16.338	19.511	21.615	24.769	27.587	30.995	33.409
18	7.015	7.906	9.390	10.865	12.857	14.440	17.338	20.601	22.760	25.989	28.869	32.346	34.805
19	7.633	8.567	10.117	11.651	13.716	15.352	18.338	21.689	23.900	27.204	30.144	33.687	36.191
20	8.260	9.237	10.851	12.443	14.578	16.266	19.337	22.775	25.038	28.412	31.410	35.020	37.566
21	8.897	9.915	11.591	13.240	15.445	17.182	20.337	23.858	26.171	29.615	32.671	36.343	38.932
22	9.542	10.600	12.338	14.041	16.314	18.101	21.337	24.939	27.301	30.813	33.924	37.659	40.289
23	10.196	11.293	13.091	14.848	17.187	19.071	22.337	26.018	28.429	32.007	35.172	38.968	41.638
24	10.856	11.992	13.848	15.659	18.062	19.943	23.337	27.096	29.553	33.196	36.415	40.270	42.980
25	11.524	12.697	14.611	16.473	18.940	20.867	24.337	28.172	30.675	34.382	37.652	41.566	44.314
26	12.198	13.409	15.379	17.292	19.820	21.792	25.336	29.246	31.795	35.563	38.885	42.856	45.642
27	12.879	14.135	16.151	18.114	20.703	22.719	26.336	30.307	32.912	36.741	40.135	44.140	46.972
28	13.565	14.847	16.928	18.939	21.588	23.647	27.336	31.361	34.027	37.916	41.401	45.429	48.297
29	14.236	15.574	17.708	19.768	22.453	24.557	28.336	32.810	35.139	39.082	42.659	46.721	49.598
30	14.953	16.306	18.493	20.599	23.364	25.468	29.336	33.550	36.250	40.256	43.773	47.962	50.892

* Table E is reprinted from Table III of Fisher's Statistical methods for research workers. Oliver & Boyd, Edinburgh, by kind permission of the author and publishers.

TABLE F. 4.—5 PER CENT (ROMAN TYPE) AND 1 PER CENT (BOLD-FACED TYPE) POINTS FOR THE DISTRIBUTION OF F

η^2	η^2 degrees of freedom (for greater variance)																								∞
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500		
1	161 4,052	200 4,999	216 5,403	225 5,625	230 5,764	234 5,869	237 5,928	239 5,981	241 6,022	242 6,056	243 6,082	244 6,106	245 6,142	246 6,169	248 6,208	249 6,234	250 6,258	251 6,286	252 6,308	253 6,323	254 6,352	254 6,361	254		
2	18.51	19.00	19.16	19.25	19.30	19.33	19.36	19.37	19.38	19.39	19.40	19.41	19.42	19.43	19.44	19.45	19.46	19.47	19.48	19.49	19.49	19.50	19.50		
3	10.13	9.55	9.28	9.12	9.01	8.94	8.88	8.84	8.81	8.78	8.76	8.74	8.71	8.69	8.66	8.64	8.62	8.60	8.58	8.57	8.56	8.54	8.53		
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.93	5.91	5.87	5.84	5.80	5.77	5.74	5.71	5.70	5.68	5.65	5.64	5.63		
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.78	4.74	4.70	4.68	4.64	4.60	4.56	4.53	4.50	4.46	4.44	4.42	4.40	4.38	4.37		
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.96	3.92	3.87	3.84	3.81	3.77	3.75	3.72	3.71	3.69	3.68		
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.63	3.60	3.57	3.52	3.49	3.44	3.41	3.38	3.34	3.32	3.29	3.28	3.25	3.24		
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.34	3.31	3.28	3.23	3.20	3.15	3.12	3.08	3.05	3.03	3.00	2.98	2.96	2.94		
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.13	3.10	3.07	3.02	2.98	2.93	2.90	2.86	2.82	2.80	2.77	2.76	2.73	2.72		
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.97	2.94	2.91	2.86	2.82	2.77	2.74	2.70	2.67	2.64	2.61	2.59	2.56	2.55		
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.86	2.82	2.79	2.74	2.70	2.65	2.61	2.57	2.53	2.50	2.47	2.45	2.42	2.41		
12	4.75	3.88	3.49	3.26	3.11	3.00	2.92	2.85	2.80	2.76	2.72	2.69	2.64	2.60	2.54	2.50	2.46	2.42	2.40	2.36	2.35	2.32	2.31		
14	4.60	3.74	3.34	3.11	2.96	2.85	2.77	2.70	2.65	2.60	2.56	2.53	2.48	2.44	2.39	2.35	2.31	2.27	2.24	2.21	2.19	2.16	2.13		
	8.86	6.51	5.56	6.03	4.59	4.46	4.28	4.14	4.03	3.94	3.86	3.80	3.70	3.62	3.51	3.43	3.34	3.26	3.21	3.14	3.06	3.02	3.00		

TABLE F.*—5 PER CENT (ROMAN TYPE) AND 1 PER CENT (BOLD-FACED TYPE) POINTS FOR THE DISTRIBUTION OF F_{∞} —(Continued)

n_2		n_1 degrees of freedom (for greater variance)																				n_1			
		1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞
17	4.45	3.59	3.20	2.96	2.81	2.70	2.62	2.55	2.50	2.45	2.41	2.38	2.33	2.29	2.23	2.19	2.15	2.11	2.08	2.04	2.02	1.99	1.97	1.96	1.96
	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.52	3.45	3.36	3.27	3.16	3.08	3.00	2.92	2.86	2.79	2.76	2.70	2.67	2.65	2.65
	20	4.35	3.49	3.10	2.87	2.71	2.60	2.52	2.45	2.40	2.35	2.31	2.28	2.23	2.18	2.12	2.08	2.04	1.99	1.96	1.92	1.90	1.87	1.85	1.84
	24	8.10	5.85	4.94	4.43	4.10	3.87	3.71	3.56	3.45	3.37	3.30	3.23	3.13	3.05	2.94	2.86	2.77	2.69	2.63	2.56	2.53	2.47	2.44	2.42
24	4.26	3.40	3.01	2.78	2.62	2.51	2.43	2.36	2.30	2.26	2.22	2.18	2.13	2.09	2.02	1.98	1.94	1.89	1.84	1.82	1.80	1.76	1.74	1.73	1.73
	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.25	3.17	3.09	3.03	2.93	2.85	2.74	2.66	2.58	2.49	2.44	2.36	2.33	2.27	2.23	2.21	2.21
	30	4.17	3.32	2.92	2.69	2.53	2.42	2.34	2.27	2.21	2.16	2.12	2.09	2.04	1.99	1.93	1.89	1.84	1.79	1.76	1.72	1.69	1.66	1.64	1.62
	40	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.06	2.98	2.90	2.84	2.74	2.66	2.56	2.47	2.38	2.29	2.24	2.16	2.13	2.07	2.03	2.01
30	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.07	2.04	2.00	1.95	1.90	1.85	1.80	1.74	1.69	1.66	1.61	1.59	1.55	1.53	1.51	1.50
	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.88	2.80	2.73	2.66	2.56	2.46	2.39	2.26	2.18	2.09	2.00	1.94	1.86	1.82	1.76	1.71	1.68
	50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.02	1.98	1.95	1.90	1.85	1.78	1.74	1.69	1.63	1.60	1.55	1.52	1.48	1.46	1.44
	70	7.17	5.06	4.20	3.72	3.41	3.18	3.02	2.88	2.78	2.70	2.62	2.56	2.46	2.39	2.26	2.13	2.00	1.90	1.82	1.74	1.69	1.62	1.56	1.53
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.01	1.97	1.93	1.89	1.84	1.79	1.72	1.67	1.62	1.57	1.53	1.47	1.45	1.40	1.37	1.35	1.35
	7.01	4.92	4.08	3.60	3.29	3.07	2.91	2.77	2.67	2.59	2.51	2.45	2.35	2.26	2.15	2.07	1.98	1.88	1.82	1.74	1.69	1.62	1.56	1.53	1.53
	100	3.94	3.09	2.70	2.46	2.30	2.19	2.10	2.03	1.97	1.92	1.88	1.85	1.79	1.75	1.68	1.63	1.57	1.51	1.48	1.42	1.39	1.34	1.30	1.28
	150	6.90	4.82	3.98	3.51	3.20	2.99	2.82	2.69	2.59	2.51	2.43	2.36	2.26	2.19	2.08	1.98	1.89	1.79	1.73	1.64	1.59	1.51	1.46	1.43
150	3.91	3.06	2.67	2.43	2.27	2.16	2.07	2.00	1.94	1.89	1.85	1.82	1.76	1.71	1.64	1.59	1.52	1.45	1.42	1.37	1.34	1.29	1.25	1.22	1.22
	6.81	4.75	3.91	3.44	3.14	2.92	2.76	2.62	2.53	2.44	2.37	2.30	2.20	2.12	2.00	1.91	1.83	1.72	1.66	1.56	1.51	1.43	1.37	1.33	1.33
	200	3.89	3.04	2.65	2.41	2.26	2.14	2.05	1.98	1.92	1.87	1.83	1.80	1.74	1.69	1.62	1.57	1.52	1.45	1.42	1.35	1.32	1.26	1.22	1.21
	400	6.76	4.71	3.88	3.41	3.11	2.90	2.73	2.60	2.50	2.41	2.34	2.28	2.17	2.09	1.97	1.88	1.79	1.69	1.62	1.53	1.48	1.39	1.33	1.28
400	3.86	3.02	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85	1.81	1.78	1.72	1.67	1.60	1.54	1.49	1.42	1.38	1.32	1.28	1.22	1.16	1.13	1.13
	6.70	4.66	3.83	3.36	3.06	2.85	2.69	2.55	2.46	2.37	2.29	2.23	2.12	2.04	1.92	1.84	1.74	1.64	1.57	1.47	1.42	1.32	1.24	1.19	1.19
	1,000	3.85	3.00	2.61	2.38	2.22	2.10	2.02	1.95	1.89	1.84	1.80	1.76	1.70	1.65	1.58	1.51	1.47	1.41	1.36	1.30	1.26	1.19	1.13	1.08
	1,000	6.66	4.62	3.80	3.34	3.04	2.82	2.66	2.53	2.43	2.34	2.26	2.20	2.09	2.01	1.89	1.81	1.71	1.61	1.54	1.44	1.38	1.28	1.19	1.11
∞	3.84	2.99	2.60	2.37	2.21	2.09	2.01	1.94	1.88	1.83	1.79	1.75	1.69	1.64	1.57	1.52	1.46	1.40	1.35	1.28	1.24	1.17	1.11	1.00	1.00
	6.54	4.60	3.78	3.32	3.03	2.80	2.64	2.51	2.41	2.32	2.24	2.18	2.07	1.99	1.87	1.79	1.69	1.59	1.52	1.41	1.36	1.25	1.15	1.00	1.00

* Reproduced from Snedecor, G. W. *Statistical methods* Ames, Iowa: Collegiate, 1937. Pp. 174-177. By permission of the author.

AUTHOR INDEX

A

Adkins, D C , 299, 299*n*.

B

Baller, W R., 166*n*.
Baxter, B., 153*n*.
Beers, F. S., 25*n*.
Bernreuter, R. G., 301
Binet, A., 284
Brown, W., 275*f*, 282, 284

C

Cantill, H , 164*n* , 170*n*.
Chesire, L , 244*n*.
Cobb, M. V., 245*n*.
Conrad, H. S., 289*n*.
Cox, H. M., 25*n*.

D

Dressel, P. L., 278*n*.
DuBois, P. H., 231, 231*n*.
Dunlap, J. W., 13, 239*n*.

E

Enlow, E. R., 13
Ezekiel, M., 13

F

Fisher, R. A., 13, 153*n*, 161, 161*n*,
209*f* , 236, 295, 325*n*.
Flanagan, J. C., 296*n*.

G

Gallup, G., 163
Galton, F., 149

Garrett, H. E., 13
Goodfellow, L. D., 160*n*.
Gray, C. T., 13
Guilford, J P , 13, 109*n*, 114*n*, 116*n* ,
146*n*, 285*n* , 288*n*, 293*n*, 296*n*,
299*n*.
Guilford, R. B., 285*n*, 288*n*
Guttman, L., 181*f*.

H

Hartson, L. D , 257*n*.
Helson, H., 146
Holzinger, K. J., 13
Hoyt, C. J., 278*n*.

J

Jenness, A. F., 196*n*.
Jorgensen, A. P , 196*n*.

K

Kelley, T. L., 13, 236
Kuder, G. F., 276*f* , 276*n*, 278*n*.
Kurtz, A K., 13

L

Laird, D. A., 284
Likert, R., 110*n*.
Lindquist, E. F., 13, 126*n*, 153, 153*n*,
210*n*, 251*n*.

M

McCall, W. A., 100
McNemar, Q., 132*n*, 141*n*.
McQuitty, J. V., 296*n*, 298, 298*n*.
Martin, G. B., 289*n*.
Mosier, C. I., 285, 296*n*, 298, 298*n*.

P

- Pearson, K , 206*f* , 223*f* , 226*f* , 229–231, 230*n* , 235–237, 242–244, 247, 253, 253*n* , 255, 262, 270, 296*f*
 Peters, C. C , 13, 131, 131*n.*, 210, 236, 236*n* , 253, 253*f.n.*

R

- Richardson, M. W , 276*f* , 276*n.*, 278*n* , 291*n* , 299*n.*
 Rulon, P J , 275*f* , 276*n.*, 280

S

- Saffir, M , 244*n*
 Seashore, C., 293
 Sletto, R. F , 295*n.*
 Snedecor, G W , 13, 147, 153*n* , 169*n* , 323*n.*, 327*n.*
 Sorenson, H , 13, 305*n* –317*n*

- Spearman, C , 275*f.*, 282, 284
 Strong, E. K , 301

T

- Thurstone, L L , 244*n* , 254, 285*n.*
 Toops, H. A , 298
 Treloar, A E , 13, 162*n*

V

- Van Voorhis, W R., 13, 131, 131*n* , 210, 236, 236*n* , 253, 253*f n*
 Votaw, D F , 13

W

- Walker, H M., 130*n* , 131
 Wallace, H. A , 323*n.*
 Weber, E H , 62
 Woodworth, R. S., 284

Y

- Yates, F., 169

SUBJECT INDEX

A

Abac, for item weight, 301
 for phi coefficient, 297
 Affective scale, 110
 Analysis of variance, 145-153, 235*f*.
 Attenuation, correction for, 287*f*
 Attributes, prediction of, 178-187
 Average, choice of, 40-44
 deviation, 59*f*.
 Averages, 28*ff*.
 correlation of, 251
 running, 23*f*.

B

Bar diagram, of distributions, 74
 Beta coefficient, 259, 265*f*.
 Biserial correlation, 237-240, 295

C

Centile, norms, 67-73
 position, 107
 profile chart, 73
 Centiles, and *z*-scores, 91
 Chi square, 167-173, 237, 246, 297
 table of, 325
 Class interval, limits of, 15*f*.
 size of, 14*f*.
 Coefficient, of alienation, 222
 of correlation, 198
 interpretation, 218-223
 significance, 323*f*.
 of determination, 223, 279
 multiple, 261
 of nondetermination, 223
 multiple, 261
 of variation, 61*f*.
 Coin tossing, probabilities in, 79
 Computation, significant figures in,
 10-12
 Correction, for attenuation, 287*f*.

Correction, for chance success, 116
 in guessed mean, 31-33
 Correlation, diagram, 192*f*
 ratio, 224, 231-236
 Criterion, for validity, 286*f*.
 Critical ratio, 298
 C-scale, 104-107
 from ranks, 108
 Cumulative, distributions, 64*ff*.
 frequencies, 64*f*.
 percentages, 66*f*.
 proportions, 66*f*.

D

Deciles, 67*f*.
 criticism of, 70*f*.
 Degrees of freedom, 130, 147, 169
 Dependent variable, 213, 256
 Design, of experiments, 153
 Difficulty, item, 114-118, 292-294
 Doolittle method, 263

E

Errors, of grouping, 253
 of measurement, 144*f*
 Extrasensory perception, 156*f*.

F

F ratio, Snedecor's, 147
 table, 326*f*.
 Factor, analysis, 285
 loading, 285
 Fiducial limits, 129*f*, 161*n*
 Forecasting efficiency, 179*f*
 Frequencies, cumulative, 64*f*.
 Frequency polygon, 17-19

G

Gallup poll, 163
 Gaussian curve, 76*ff*
 Goodness of fit, to normal distribution,
 173

H

Histogram, 20*f*.
Homoscedasticity, 224

I

Independent variable, 213, 256
Index, correlations, 252*f*
 of forecasting efficiency, 222*f*., 262,
 289
 of reliability, 278
Internal consistency, of items, 294*f*.
Intravariability, 98*f*.
Item analysis, 292-302

J

Judgments, scaling of, 110*f*.

L

Least squares, principle of, 188
Linearity, test of, 236*f*.
L-method, 299

M

Maximum likelihood, 178*f*.
Mean, arithmetic, 29-33
 geometric, 28
 guessed, 31-33
 harmonic, 28
 reliability of, 125*ff*.
Measurement, advantages of, 3-5
 educational, 2*f*.
 errors of, 144*f*.
 psychological, 1*ff*.
 and statistics, 1
Median, 34-39
 reliability of, 132
Mode, crude, 39
 estimated, 39*f*.
Multiple correlation, 256*ff*.
 coefficient, 266*f*.
Multiple regression equation, 258-261

N

Normal curve, assumption, 76*f*.
 best-fitting, 81, 83

Normal curve, equation, 80
 tables of, 318-322

Normalized distributions, 95
 hypothesis of, 173
Null hypothesis, 156-167
Numbers, approximate, 8
 in measurement, 7*f*.
 rounding of, 8*f*.

O

Ogive, 67
 centile norms from, 70

P

Partial correlation, 268-271
Pearson *r*, formula, 202-204, 206
Phi coefficient, 245-248
Prediction, accuracy of, 177*f*., 186*f*.
 from equations, 215*f*.
 of measurements, 188-196
 multiple, 260
 types of, 176*f*.
Primary abilities, 285
Probability, curve, 78
Probable error, 55*n*.
Product-moment correlation, 202
Profile, *C*-score, 106
 graphic chart, 72

Q

Quartile, standard error of, 133
Quartiles, 49
 graphic determination of, 65*f*.

R

Range, middle 80 per cent, 48*f*
 semi-interquartile, 49-51
 and standard deviation, 56
 total, 47*f*.
Rank difference, correlation, 227-230
Rank order, 107-110
 as measurement, 5
Ratings, reliability of, 284
 transformation of, 118
Rational equivalence, 276*ff*.

Regression, coefficient, 212-214, 259
 equation, 211-214
 line, 195
 weight, multiple, 266f
 Relativity, of coefficients of correlation, 248
 Reliability, of averages, 125-132
 coefficient, 219
 of differences, 135-144
 and length of test, 282
 of proportions, 133f.
 and range of measurement, 281
 of the standard deviation, 132
 of test scores, 273-284
 and time-limit, 283f.
 Rho, coefficient, 227
 Richardson-Kuder formula, 277
 Rulon formula, 276

S

Sampling, biased, 77
 Scaling, of observations, 95f.
 Scatter diagram, 205f
 Significant figures, 10-12
 Skewness, 23, 41, 50f.
 causes of, 77
 Small samples, theory of, 129ff.
 Smoothing, of distributions, 23f.
 Spearman-Brown formula, 275, 282
 Spurious correlations, 251-253
 Square roots, table of, 305-317
 Squares, table of, 305-317
 Standard deviation, 51-59
 interpretation of, 53
 Standard error, of biserial r , 239
 of correlation ratio, 235
 defined, 127
 of a difference, 135-144
 of estimate, 190f, 194, 216f., 235, 262
 of a frequency, 134
 of the mean, 125ff.
 of median, 132
 of an obtained score, 279
 of a percentage, 134
 of a proportion, 133
 of Q , 133

Standard error, of r , 209
 of a regression coefficient, 218
 of the standard deviation, 132
 of a tetrachoric r , 243f.
 Standard measurements, 82
 Standard scores, 82
 disadvantages of, 99
 Student's distribution, 130
 Student's ratio, 130
 Successive intervals, method of, 110

T

Test scores, as measurements, 6f.
 Tetrachoric correlation, 240-245, 296
 Transformation, of distributions, 118-122
 True mean, defined, 125
 t -ratio, defined, 130
 table for, 323f
 T -scale, 99-104
 disadvantages of, 104
 from ranks, 108

V

Validity, coefficient, 219
 and difficulty of items, 293f.
 of test batteries, 290f.
 of test items, 294-298
 and test length, 289
 of test scores, 284-292
 of tests, 273
 Variability, and correlation, 248f.
 relative, 61f
 Variance, analysis of, 145-153
 defined, 145
 of test item, 277
 Variation, coefficient of, 61f.

W

Weber's law, 62
 Weight, for test items, 298-302
 Weighting, of tests, 291

Y

Yates' correction for continuity, 169